THE UNIVERSITY OF CHICAGO

ESSAYS ON ECONOMETRICS AND INDUSTRIAL ORGANIZATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY
JONAS LIEBER

CHICAGO, ILLINOIS
JUNE 2023

Contents

Preface

List of Figures

List of Tables

Acknowledgements

First, I am grateful to my advisors, Alexander Torgovitsky, Ali Hortaçsu, Azeem Shaikh and Max Tabord-Meehan, for their invaluable guidance and support. They have been outstanding in evaluating early stage ideas, providing feedback the process of developing some of these ideas, and in giving advice for and supporting me on the job market.

Further, I would like to thank Ali Hortaçsu, Áureo de Paula and Julien Monardo for the opportunity to collaborate on what became the second chapter of this dissertation. The differences in our background and research focus have led to questions, thoughts, and comments which have often led me to insights.

Next, I would like to thank Thomas Wiemann for his contribution to the work that became the third chapter of this dissertation. Here, the similarity of our backgrounds has been helpful in the design of the algorithm and the development of the theoretical results.

Although our joint work has not been completed in time to be included in this dissertation, I would also like to thank Alexander Torgovitsky, Pietro Tebaldi and Hyong-gu Hwang for the opportunity to collaborate on a project on semiparametric demand estimation that made me grow as a researcher.

When my office at The University of Chicago closed due to the pandemic, Hannes Ullrich and Tomaso Duso invited me to the DIW Berlin where I could continue my research in a wonderful and productive atmosphere for several months. To Hannes Ullrich, I owe a special thanks for the trust he had in me to be a researcher even before I started this doctoral program. I hope that our joint work that has not been completed in time for the submission of this dissertation will be com-

pleted soon.

For helpful comments, I thank Sté phane Bonhomme, Eyo Herstad, Evgueni Kivman, Julien Monardo, Jesper Riis-Vestergaard Sørensen, Thomas Wiemann, and participants of the Econometrics Advising Group, the Econometrics Students Group, and the Industrial Organization Lunch at The University of Chicago.

Finally, I would like to thank Sergei Bazylik, Arjun Gopinath, Agustin Gutierrez, Takuma Habu, Eyo Herstad, Thomas Hierons, Sota Ichiba, Rafael Jimenez, Esperanza Johnson, Noemi Nocera, Ricardo Quineche, Francesco Ruggieri, Harshil Sahai, Myungkou Shin, Mateusz Stalinski, Francisco Del Villar, Ana Vasilj, Thomas Wiemann and the "Squirrels of Hyde Park" for their friendship and support.

Finally, I would like to thank my family for their unconditional support throughout a lengthy formal education process that now finally comes to an end.

Legal Notice on Chapter 1

The results in this paper reflect the researcher's own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Abstract

This dissertation is comprised of three chapters. The unifying theme of this dissertation is the estimation of demand. The first chapter proposes a method to estimate concentration parameters for bandit algorithms. It can be applied to experimenting with prices when demand is initially unknown. The second chapter develops a method to estimate nesting structures in demand models. The third chapter considers the estimation of demand when there are finitely many consumers, specifically the zero-valued market shares.

Chapter 1

Estimating Concentration Parameters for Bandit Algorithms

1.1 Introduction

Following the repeal of prohibition in 1933, the State of Washington created a monopoly for selling liquor. This monopoly was upheld for almost 80 years until voters adopted an initiative that led to the privatization of liquor sales starting June 1st, 2012. Suddenly, vendors were free to sell liquor and to set prices. To set profit-maximizing prices, firms would have to know the demand curve. There is anecdotal evidence that vendors did not know the demand curve. Alan Johnson, CEO of BevMo!, a Californian company specializing in the sale of alcoholic beverages that entered the liquor markets in Washington, said in March 2012:

"I sure don't know what we'll charge the consumer.
There is going to be a lot of scrambling."[1]

In this situation, a firm can learn the demand curve by simply setting a price and then observing

[1] Cited according to Huang et al. (2022) as the original source Gregutt, Paul, "BevMo Ramps It Up in Washington State" (April 30, 2012) is no longer available.

demand. But learning the demand curve is only a means to an end, profit maximization. This problem can be modeled as a multi-armed bandit (Rothschild, 1974). In multi-armed bandits models, agents repeatedly choose actions with uncertain rewards. While learning about the distribution of rewards associated with each action, agents aim to maximize expected utility or profits. Many dynamic programming problems can be written as bandit models. Examples include firms maximizing profits by experimenting with prices or advertisement (Rothschild, 1974; Schwartz et al., 2017; Misra et al., 2019; Waisman et al., 2019), randomized control trials maximizing outcomes by evaluating alternative treatments (Berry, 2006; FDA, 2018; Kasy and Sautmann, 2021), consumers maximizing utility by trying experience goods (Hotz and Miller, 1993; Erdem and Keane, 1996; Crawford and Shum, 2005), or workers maximizing income by choosing an industry or a job (Keane and Wolpin, 1997; Jovanovic, 1979; Miller, 1984).

A popular bandit algorithm is the upper confidence bound (UCB) algorithm. The UCB algorithm advocates for optimism in the face of uncertainty. Formally, it constructs finite-sample confidence intervals (CIs) for the mean reward associated with each action. The UCB algorithm then recommends the action with the highest upper confidence bound, i.e., the action that gives the most reason for optimism. The construction of these finite-sample CIs uses Hoeffding's inequality, a probabilistic bound on how much the average can deviate from the mean.

A caveat in applying Hoeffding's inequality is that it is not fully data-driven. It involves concentration parameters that govern the tail. These parameters resemble the variance in the CLT, where consistent estimates of the variance exist and can, by Slutsky's Lemma, be used for inference. So far, there has been no analog of these results for inference with Hoeffding's inequality.

In this paper, I propose two methods for estimating the concentration parameters which appear in Hoeffding's inequality. I establish that asymptotic inference with estimated parameters is valid,

i.e., confidence intervals have at least the nominal coverage under mild conditions. Under stronger assumptions, I show that the estimated parameters are asymptotically optimal, i.e., they yield the narrowest possible confidence interval in the class of valid confidence intervals based on Hoeffding's inequality. In contrast, I show the impossibility of finite-sample inference with estimated parameters without further assumptions. With an additional anti-concentration assumption, I show that finite-sample bounds with estimated parameters are feasible.

In the statistical literature on bandits, concentration parameters appearing in Hoeffding's inequality are assumed to be known (Lattimore and Szepesvári, 2020). In practice, however, these concentration parameters are either unknown or estimated from above using support bounds. These support bounds often lead to inflated concentration parameters. As a result, upper confidence bounds are unnecessarily high, leading to an overexploration of sub-optimal options in the UCB algorithm. I leverage the methods derived in this paper to adapt the UCB algorithm to settings where the concentration parameters are unknown. Under an anti-concentration assumption, I derive a finite-sample bound for the regret of the adapted UCB algorithm that is close to the finite-sample regret bound of the asymptotically optimal UCB algorithm with known parameters.

Subsequently, I apply these methods to the optimal price experimentation problem of firms entering the liquor market in Washington state in 2012 where entering firms experimented to learn the shape of the demand curve (Huang et al., 2022). I empirically compare the price experimentation implemented by store managers with UCB algorithms with concentration parameters from support bounds and the proposed estimators. I find that the UCB algorithm with parameters based on support bounds would have increased profits by 1% . The UCB algorithm with estimated concentration parameters achieves a 26% profit increase, close to the 30% increase of the infeasible UCB algorithm with optimal parameters.

My theoretical results can also be applied to non-standard inference problems that arise in partial identification and machine learning. Inference problems in partial identification and machine learning are often difficult because they involve non-Gaussian limiting distributions. One example is when the true parameter lies on the boundary of the parameter space. In the case of the LASSO (Tibshirani, 1996), which has been used to approximate optimal instruments (Belloni et al., 2012; Gilchrist and Sands, 2016) and for inference on treatment effects in the presence of many controls (Belloni et al., 2013), Fu and Knight (2000) show that the limiting distribution can have probability mass on zero when the true coefficient is zero. A second example are non-differentiable functions of the mean. In the case of linear programs, which are widely used in partial identification of, e.g., counterfactual demand (Tebaldi et al., 2019), policy-relevant treatment effects (Mogstad et al., 2018), and peer effects (Herstad, 2022), this occurs when (some) parameters of the linear program are estimated (Mangasarian and Shiau, 1987). Delta method and the bootstrap are invalid without differentiability Shapiro (1990); Fang and Santos (2019)). In this paper, I show that Hoeffding's inequality can be used to derive conservative inference in these settings even when there are non-differentiabilities or parameters lie on the boundary of the parameter space.

I propose a finite-sample inference to linear programs, allowing for all parameters to be estimated. I show how the problem can be treated as a quadratically constrained quadratic program and derive sufficient conditions under which this program is convex. This allows the number of parameters to grow almost exponentially in the sample size. I illustrate the proposed method in a simulation study based on the nonparametric demand estimation strategy proposed by Tebaldi et al. (2019).

Using Hoeffding's inequality for inference has a price. First, inference based on Hoeffding's inequality tends to be conservative, i.e., coverage tends to be above the nominal level. For example, a 95% confidence interval of a standard normal based on Hoeffding's inequality is about 39% larger than it would have to be, its coverage is 99.35%. Second, the necessary tail assumptions are

stronger than the tail assumptions needed for the CLT. Hence, inference based on the CLT needs weaker assumptions and yields narrower confidence intervals. As discussed above, the strengths of inference based on Hoeffding's inequality are finite-sample validity, that parameters do not have to be in the interior, and that differentiability is not required when passing from inference on the mean to inference on a function of the mean.

The rest of this paper is organized as follows. Section 1.1.1 discusses the related literature. In section 1.2, I present empirical settings that motivate this paper. I develop estimators for sub-Gaussian concentration parameters in section 1.3 and study inference with these estimated parameters in section 1.4. The motivating examples are revisited in section 1.5. The empirical application to liquor sales in Washington is in section 1.6. Section 3.6 concludes.

1.1.1 Related Literature

This paper relates to several strands of literature.

First, it relates to the econometric literature on finite-sample inference on the mean. Bahadur and Savage (1956) show the non-existence of reasonable finite-sample confidence intervals for the mean as long as the class of distributions considered is sufficiently large. When the class of distributions is restricted to those with *known* support bounds, Hoeffding (1963) establishes valid finite-sample confidence intervals whose length decreases at the optimal rate $\sqrt{n}$ of the CLT but feature a sub-optimal constant compared to confidence intervals based on the CLT. When the class of distributions is restricted to those with *known* support bounds, Romano and Wolf (2000) build on Anderson (1969) to propose confidence intervals with uniform finite-sample validity and asymptotically optimal length. However, these confidence intervals are hard to compute as they involve infinite-dimensional optimization over the class of all distributions. In this paper, I show that finite-sample inference based on Hoeffding's inequality with estimated concentration parameters

is impossible without further assumptions, echoing Bahadur and Savage (1956). When inference based on Hoeffding's inequality with estimated parameters is possible, the confidence intervals are asymptotically wider than those derived with the CLT. This is the price for simplicity of computation, finite-sample validity, and the ability to sidestep (directional) differentiability requirements in asymptotic inference.

Second, this paper relates to the econometric literature on inference on functions of the mean and the literature on moment inequalities. Shapiro (1990) and Fang and Santos (2019) show that (directional) differentiability is required to pass from inference on the mean to inference on a function of the mean using the delta method or the bootstrap. Many interesting inference problems do not satisfy this differentiability requirement, lack asymptotic normality and are hence considered nonstandard. Examples include linear programs (Mangasarian and Shiau, 1987) and the LASSO (Fu and Knight, 2000). Using inference based on concentration inequalities does not require differentiability or even continuity. I contribute to this literature by deriving conditions under which inference based on Hoeffding's inequality can be used with estimated concentration parameters.

Third, this paper relates to the idea of using estimated concentration parameters for inference. For distributions with known support bounds, Maurer and Pontil (2009) derive a Bernstein concentration inequality with estimated variances. This result has seen applications in the bandit literature (Audibert et al., 2007) and empirical risk minimization (Shivaswamy and Jebara, 2010). For example, (Shivaswamy and Jebara, 2010) claim that Bernstein concentration inequalities yield tighter confidence intervals than Hoeffding's inequality when the variances are small. This is only true when support bounds are used to derive the sub-Gaussian paramters appearing in Hoeffding's inequality. The estimators for concentration parameters proposed in this paper have the same effect when used for inference with Hoeffding's inequality. In general, Hoeffding's inequality leads to stronger concentration than Bernstein's inequality (Vershynin, 2018, Section 2.8). It is useful

to highlight that while (Shivaswamy and Jebara, 2010) derives finite-sample concentration bounds for estimated variances using a support bound, this paper uses anti-concentration assumption for finite-sample bounds.

Fourth, this paper relates to the literature on multi-armed bandits. Specifically, it builds on the idea of estimating dispersion parameters of reward distributions on the fly. For the special case of bandits whose reward distributions are exactly Gaussian, Auer et al. (2002) propose and study the UCB-Normal algorithm that estimates the unknown variances on the fly. For general reward distributions, Auer et al. (2002) proposed the experimental UCB1-Tuned algorithm with estimated variances which they found to perform well in simulations. Audibert et al. (2007) leverage the Bernstein concentration inequality with known bounded support derived by Maurer and Pontil (2009) to propose and study the UCBV algorithm that uses estimated variances. In bandits with different variances per arm, this reduces the regret by experimenting less on arms that appear to have low variances. Mukherjee et al. (2018) refine EUCBV by combining UCBV with the arm-elimination strategy of UCB-Improved developed by Auer and Ortner (2010). Honda and Takemura (2011) propose the DMED algorithm motivated by a Bayesian perspective and establish its instance-specific asymptotic optimality. All these bandit algorithms require known support bounds for reward distributions. My refinement of the UCB algorithm does not require that the support is bounded nor that the support bounds are known. Under an additional anti-concentration property, I establish a finite-sample regret bound of the UCB algorithm with estimated parameters that is close to the regret bound of the UCB algorithm with known parameters. In my empirical application, I find that DMED performs comparably to UCB with concentration parameters based on support bounds. Using estimated concentration parameters improves the performance in this example for a fixed horizon.

Fifth, this paper adds to the literature on inference to linear programs with estimated parame-

ters. Fang et al. (2022) propose a method for inference when all coefficients are known but the right hand side in the constraint is estimated. Hsieh et al. (2022) consider using the optimality conditions of linear (and quadratic) programs as moment inequalities to leverage methods from the literature on inference with moment inequalities. This method is computationally challenging, particularly when there are many parameters and constraints. In the most closely related paper in this literature, Syrgkanis et al. (2021) study inference for a linear program that arises in non-parametric inference on auctions using finite-sample inference. In their linear program, the coefficients in the objective are estimated as means of sub-Gaussians with known sub-Gaussian parameters. Syrgkanis et al. (2021) note that using Hoeffding's inequality is computationally attractive. While the methods proposed in this paper would solve a quadratic program for inference, Syrgkanis et al. (2021) use the particular structure of their linear program to derive a inference method that only relies on solving a *linear* program. My paper adds to this literature by covering the general case in which all parameters are estimated, sub-Gaussian parameters are unknown, and computational complexity due to many variable or constraints is a first-order concern.

Sixth, this paper contributes to the empirical literature on firm behavior, learning and deviations from optimality. Rothschild (1974) showed that the optimal price experimentation strategy of a Bayesian monopolist leads to a positive probability of committing to a price that does not maximize profits. DellaVigna and Gentzkow (2019) show that retail chains forgo substantial profits by adopting uniform pricing policies that are not adapted to geographically heterogeneous demand conditions. Hortaçsu et al. (2021) find that even a large US airline using sophisticated methods to estimate demand is subject to certain biases and hence sets sub-optimal prices. In contrast, Huang et al. (2022) find that firms entering the liquor market in Washington State are successful in learning complex features of demand and in setting optimal prices within 2.5 years of entry. Studying the same setting, I find that firms could have considerably increased their profits in the 2.5 years of experimentation. According to my demand estimates, firms did not succeed in finding the optimal

prices.

1.2 Applications

The purpose of this section is to illustrate empirical settings which serve as motivating examples for inference based on concentration inequalities. First, I introduce multi-armed bandits as a model of optimal experimentation for a monopolist facing an unknown demand curve. Second, I consider applications to partial identification, specifically inference to linear programs with estimated parameters, and inference to partially identified parameters that are solutions to a class of optimization problems considered by Horowitz and Lee (2022). Third, I consider applications to machine learning. Specifically, I discuss how convergence rates can be used for inference, and the role of Hoeffding's inequality for empirical risk minimization and the selection of the batch size in the stochastic gradient descent method that is used in artificial neural networks.

Example 1. *Multi-armed bandits.* Consider a monopolist facing an unknown demand curve (Rothschild, 1974). The monopolist's objective is to maximize the expected sum of profits arising over T periods. Each period, the monopolist can choose a price from a finite set of prices $\{p_{\text{grid},1}, \ldots, p_{\text{grid},L}\}$ and observe the associated profit. For each posted price, the demand follows a Binomial distribution where the number of trials is given by the number of consumers $n < \infty$ and the success probability is given by the aggregated choice probability $s(p)$. Given a constant marginal cost C, the monopolist aims to maximize

$$\sum_{t=1}^{T} \mathbb{E}[\Pi(p_t)] = \sum_{t=1}^{T} \mathbb{E}\left[(p_t - C)\text{Bin}(n, s(p_t))\right] = n \sum_{t=1}^{T} (p_t - C)s(p_t). \tag{1.1}$$

by choosing the sequence (p_t) in a possibly data-driven way. For example, this allows p_{10}, the price charged in period 10, to depend on the realized demand in period 1-9. While the monopolist knows the marginal cost C and the number of consumers n in the market, the monopolist does not

know $s(\cdot)$. Because the number of consumers is finite, the monopoly only receives a noisy signal of the true demand at the posted price.

A widely used algorithm to maximize the sum of expected profits (1.1) is the UCB algorithm (Lai, 1987; Agrawal, 1995; Auer et al., 2002). The UCB algorithm requires the sub-Gaussian confidence parameters $K_1, \ldots, K_L$ of the profits for each price $p_{\text{grid},1}, \ldots, p_{\text{grid},L}$. After trying every arm once, the UCB algorithm computes an upper confidence bound based on Hoeffding's inequality with known sub-Gaussian concentration parameter (Theorem 11) with a time-dependent coverage level. Then the UCB algorithm recommends the price with the highest upper confidence bound, updates the finite sample confidence interval of the chosen option. This is formalized in Algorithm 1.

If a firm wanted to implement the UCB algorithm in practice, it would have to specify the concentration parameters $K_1, \ldots, K_L$ of the profit for each price $p_{\text{grid},1}, \ldots, p_{\text{grid},L}$. The concentration parameter K_l governs the tail of the distribution of profits of charging price $p_{\text{grid},l}$. The reason why the monopolist engages in price experimentation is that these distributions are unknown. Hence, it seems unreasonable that the monopolist would know the optimal concentration parameters of the unknown distributions. The purpose of this paper is to derive a method to estimate the concentration parameters $K_1, \ldots, K_L$ while experimenting.

Many other interesting experimentation problems can be framed as bandit problems. Examples include evaluating treatments while attempting to maximize outcomes in clinical trials (Kasy and Sautmann, 2021), learning about experience goods while maximizing utility (Crawford and Shum, 2005; Hotz and Miller, 1993), learning about the match value of occupations while maximizing income (Jovanovic, 1979; Miller, 1984), learning about the promise of research projects while maximizing research output Weitzman (1979). □

Example 2. *Inference for linear programs with estimated parameters.* Consider the linear program

$$\min_{x \in \mathbb{R}^p} c'x \qquad \text{s.t.} \qquad Ax \leqslant b \tag{1.2}$$

with constraint matrix $A \in \mathbb{R}^{m \times p}$, constraint vector $b \in \mathbb{R}^m$ and objective vector $c \in \mathbb{R}^p$. The question is how to perform inference on the optimal value $c'x^*$ or the optimizer x^* of this linear program when A, b and c are expectations that are estimated.

There are many empirical settings in which this problem arises. Tebaldi et al. (2019) propose a method to compute sharp bounds on counterfactual demand without parametric assumption on latent utility draws. When prices are exogenous, c and A are observed while b is the vector of estimated market shares. Further examples include dynamic discrete choice (Nevo et al., 2016) and testing identifying assumptions in the treatment effect literature Angrist and Imbens (1995). See Fang et al. (2022) for an in-depth discussion of these and further examples. In this paper, I show how conservative inference for the linear program (1.2) can be conducted based on Hoeffding's inequality and derive sufficient conditions for computational tractability even when the number of variables and (in)equality constraints is very large. Recall that Hoeffding's inequality involves concentration parameters. The proposed method for inference can be combined with my results on using estimated concentration parameters for inference to allow for inference that does not rely on knowing concentration parameters a priori.

Inference for linear programs with estimated parameters is hard because linear programs are not (directionally) differentiable in parameters (Mangasarian and Shiau, 1987). However, the step from asymptotic inference on the mean of the parameters to asymptotic inference on a function of the mean typically requires differentiability, e.g., for the delta method or the bootstrap (Shapiro (1990); Fang and Santos (2019)). In contrast, no such condition is required for finite-sample in-

ference. As a result, finite-sample inference is available when asymptotic inference is elusive, particularly when all parameters are estimated.

I propose to derive uniform finite-sample confidence sets for A, b and c with Hoeffding's inequality. Denote these confidence sets by C_A, C_b, C_c. For inference on the optimal value the linear program (1.2), I then propose to solve

$$\min_{\substack{x\in\mathbb{R}^p \\ \tilde{c},\tilde{A},\tilde{b}}} c'x \qquad \text{s.t.} \qquad \begin{cases} \tilde{A}x \leqslant \tilde{b}, \\ \tilde{A} \in C_A, \\ \tilde{b} \in C_b, \\ \tilde{c} \in C_c. \end{cases} \tag{1.3}$$

While any method to derive finite-sample confidence intervals can be used to derive C_A, C_b, C_c, Hoeffding's inequality has the advantage that (1.3) turns out to be a convex program in relevant cases. This is important for computational tractability of (1.3) as linear programs are often used with many variables and constraints. □

Example 3. *Partial Identification.* Horowitz and Lee (2022) study inference for partially identified parameters that are solutions to a class of optimization problems. Applications include nonparametrically estimated labor supply effects of tax or welfare reforms using grouped data (Blundell et al., 1998; Kline and Tartari, 2016), entry games with multiple equilibria (Ciliberto and Tamer, 2009), and estimation under shape constraints (Freyberger and Horowitz, 2015; Horowitz and Lee, 2017) as applied to estimating the effect of childbearing on labor supply (Angrist and Evans, 1998). Two of three methods developed by Horowitz and Lee (2022) build on sub-Gaussian tail assumptions with known sub-Gaussian concentration parameters. The methods proposed in this paper to estimate these concentration parameters can be used to generalize the methods developed by Horowitz and Lee (2022) to the case of unknown concentration parameters. □

Example 4. *Machine Learning Inference via Rates.* Inference for many machine learning estimators is often difficult, see, for instance, the discussion on the LASSO estimator (Tibshirani, 1996) above. In contrast, rates of convergence are often available, e.g., for the LASSO (Bickel et al., 2009; Hastie et al., 2015), Non-Negative Least Squares (Meinshausen, 2013; Slawski and Hein, 2011, 2013) or its extensions with endogeneity (Zhu, 2018; Hortaçsu et al., 2022). These convergence rates build on Gaussian or sub-Gaussian tail assumptions to control the large number of explanatory variables. If the sub-Gaussian concentration parameters and some estimator-specific constants were known, the rate results could be used for inference. For example, non-negative two-stage least squares, used by Hortaçsu et al. (2022) to estimate nesting structures for demand models, features estimator-specific constants that are relatively simple to compute. The methods proposed in this paper to estimate sub-Gaussian concentration parameters allow inference for this high-dimensional estimator when the sub-Gaussian parameters are unknown. □

Example 5. *Empirical Risk Minimization.* Empirical risk minimization (ERM) is a paradigm in machine learning and statistics (Vapnik, 1991). ERM is used, for instance, in support vector machines (Vapnik, 1999), in boosting algorithms like adaBoost (Freund and Schapire, 1997), in decision trees (Quinlan, 1986; Breiman et al., 1984), in random forests (Breiman, 2001; Ho, 1995), and in artificial neural networks (Rosenblatt, 1958). All these methods aim to minimize an expected loss function L, i.e.,

$$\min_{h \in \mathcal{H}} R(h) = \mathbb{E}\big[L(h(X), Y)\big],$$

where h is the "hypothesis", i.e., the function we would like to learn out of a class of hypothesis $\mathcal{H}$, X are some explanatory variables, Y is some outcome and $L \geqslant 0$ is a loss function. The distribution with respect to which the expectation of the loss function is taken is unknown. Instead

of minimizing the population risk R, ERM minimizes the (training) sample risk, i.e.,

$$\min_{h \in \mathcal{H}} \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(X_i), Y_i)].$$

If $L(h(X_i), Y_i)$ is sub-Gaussian with known sub-Gaussian parameter and $\mathcal{H}$ is finite (imagine that we select one out of finitely many models), then Hoeffding's inequality can be used to derive a probabilistic bound for the "estimation error"

$$\min_{h \in \mathcal{H}} \hat{R}(h) - \min_{h \in \mathcal{H}} R(h).$$

Such a probabilistic bound can be used to find an upper bound for the number of observations required to bound the estimation error to a desired level. Extensions to infinite-dimensional models, i.e., when $|\mathcal{H}| = \infty$, follow the same logic with McDiarmid's inequality, which requires the sub-Gaussian concentration parameter of the l-th centered conditional version of the loss function. □

Example 6. *Stochastic Gradient Descent for Artificial Neural Networks.* The application to empirical risk minimization is relevant in the context of artificial neural networks (ANNs). ANNs achieve highly competitive performance in a variety of contexts, in particular to the estimation of heterogeneous treatment effects which can be used to address optimal advertisement targeting problems (Hitsch and Misra, 2018; Farrell et al., 2021). Training an ANN is often achieved via gradient descent. Even with the backpropagation algorithm, computing the derivatives of $L(h(X_i), Y_i)$ for each value of i with respect to all parameters of h (weights and biases) is challenging when there are many observations (Goodfellow et al., 2016).

To reduce the computational complexity of this optimization, stochastic gradient descent (SGD) is commonly used. In each iteration of SGD, a "mini-batch" of $k < n$ observations is drawn randomly and then the gradient of the expected loss function is only evaluated for these randomly

selected points. Choosing of the batch size k is difficult as it involves a trade-off between computational feasibility (it is easier to perform an iteration of SGD when the batch size is smaller) and convergence properties (a SGD step is more likely to decrease the empirical risk when the batch size is larger) (Bengio, 2012; Goodfellow et al., 2016). In applications, the sub-Gaussian parameter of $L(X_i, Y_i)$ may not be known. The methods proposed in this paper to estimate sub-Gaussian parameters can inform practitioners choosing the batch size in training ANNs. □

1.3 Estimating sub-Gaussian parameters

1.3.1 Definition of sub-Gaussian parameters

Two notions of sub-Gaussianity are relevant for this paper. One notion is based on a tail bound which ensures that a sub-Gaussian random variable is concentrated around its mean. A second notion of sub-Gaussianity is based on the moment-generating function (MGF), which turns out to be useful to study sums of independent random variables.

Definition 7. A real-valued random variable X is

- tail-sub-Gaussian with parameter $K > 0$ if for all $t > 0$,

$$\mathbb{P}\left[|X| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{K^2}\right), \tag{1.4}$$

- MGF-sub-Gaussian with parameter $K > 0$ if it has mean zero and for all $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leqslant \exp\left(K^2\lambda^2\right). \tag{1.5}$$

Intuitively, a random variable is tail-sub-Gaussian (MGF-sub-Gaussian) if it's tail (MGF) is dominated by the tail (MGF) of a Gaussian random variable.[2] Fortunately, these two notions of sub-

[2]See section 1.9.1.1 for details regarding the relation to Gaussians.

Gaussianity are equivalent up to a constant as is shown in section 1.4.1. Let me consider some examples of sub-Gaussians and their parameters.

Remark 8. Let X be a real-valued random variable.

1. Let $\underline{B}, \overline{B} > 0$ be such that $-\underline{B} \leqslant X \leqslant \overline{B}$. Then X is

   (a) tail-sub-Gaussian with parameter at most $\max\{\underline{B}, \overline{B}\}/\sqrt{\log(2)}$,

   (b) MGF-sub-Gaussian with parameter at most $(\overline{B} + \underline{B})/\sqrt{8}$.

2. Let $X \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma \geqslant 0$. Then X is MGF-sub-Gaussian with parameter $\sigma/\sqrt{2}$.

Remark 8 covers a wide class of distributions. An example of a bounded random variable is the profit of a firm when there are finitely many consumers, see Example 1. If $X = (p - c)s(p)n$ where p is the price, $c \leqslant p$ is the cost, $s(p)$ is a theoretical choice probability and n is the number of consumers, then X is bounded from below by 0 and from above by $(p - mc)n$. Remark 8 also shows that all Gaussians are sub-Gaussians and that the MGF-sub-Gaussian parameter and the standard deviation are identical up to a constant factor that is due to the parametrization of (1.5). As the following observation shows, there are more similarities between sub-Gaussian parameters and standard deviations.

Remark 9. If X is tail-sub-Gaussian (MGF-sub-Gaussian) with parameter K, then for any $\gamma \in \mathbb{R}$, γX is tail-sub-Gaussian (MGF-sub-Gaussian) with parameter $|\gamma| K$.

The following remark formalizes an important difference between Gaussians and Sub-Gaussians.

Remark 10. If X is tail-sub-Gaussian (MGF-sub-Gaussian) with parameter K, then for any $\kappa \geqslant 0$, X is tail-sub-Gaussian (MGF-sub-Gaussian) with parameter $K + \kappa$.

In other words, it is not true that $\mathcal{N}(0, 1)$ is also $\mathcal{N}(0, 1 + 1)$. But a sub-Gaussian with parameter 1 is also sub-Gaussian with parameter $1 + 1$. Remark 10 also implies that for every tail-sub-Gaussian (MGF-sub-Gaussian) random variable X, there is a smallest tail-sub-Gaussian (MGF-sub-Gaussian) parameter. As we will see later, smaller sub-Gaussian constants lead to narrower

confidence intervals. For sharpness of inference, it is therefore of interest to estimate the smallest tail-sub-Gaussian (MGF-sub-Gaussian) parameter.

1.3.2 Tail-sub-Gaussian parameter

In this section, I propose a method to estimate tail-sub-Gaussian parameters, starting with observed random variables. Estimated random variables, e.g., residuals in a regression, are discussed in appendix section 1.9.3.

Consider a real-valued random variable X that is tail sub-Gaussian with parameter K. Observe that X is then also $K+\varepsilon$ tail-sub-Gaussian for any $\varepsilon > 0$. The first objective is to characterize the smallest such tail-sub-Gaussian parameter of X since it allows for the least conservative inference.

$$K^*_{\text{tail}} = \min_{K \geqslant 0} K \quad \text{s.t.} \quad \forall t \geqslant 0,\ \ 1 - F_{|X|}(t) \leqslant 2\exp\left(-\frac{t^2}{K^2}\right) \tag{1.6}$$

$$= \sup_{t \in \mathbb{R}_+} \frac{t}{\sqrt{\log\left(\frac{2}{1-F_{|X|}(t)}\right)}} \tag{1.7}$$

$$= \sup_{p \in [0,1]} \frac{F^{-1}_{|X|}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}}. \tag{1.8}$$

1.3.2.1 Observed Realizations

Suppose we observe n i.i.d realizations of X, $X_1, \ldots, X_n$. To estimate K^* defined in (1.6),(1.7) and (1.8), a natural approach is to replace the unknown cdf of $|X|$ with an empirical counterpart. The empirical cdf $\hat{F}_{|X|}(t) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,|X_i|]}(t)$ is a natural estimate of the true cdf. This motivates to the following estimator

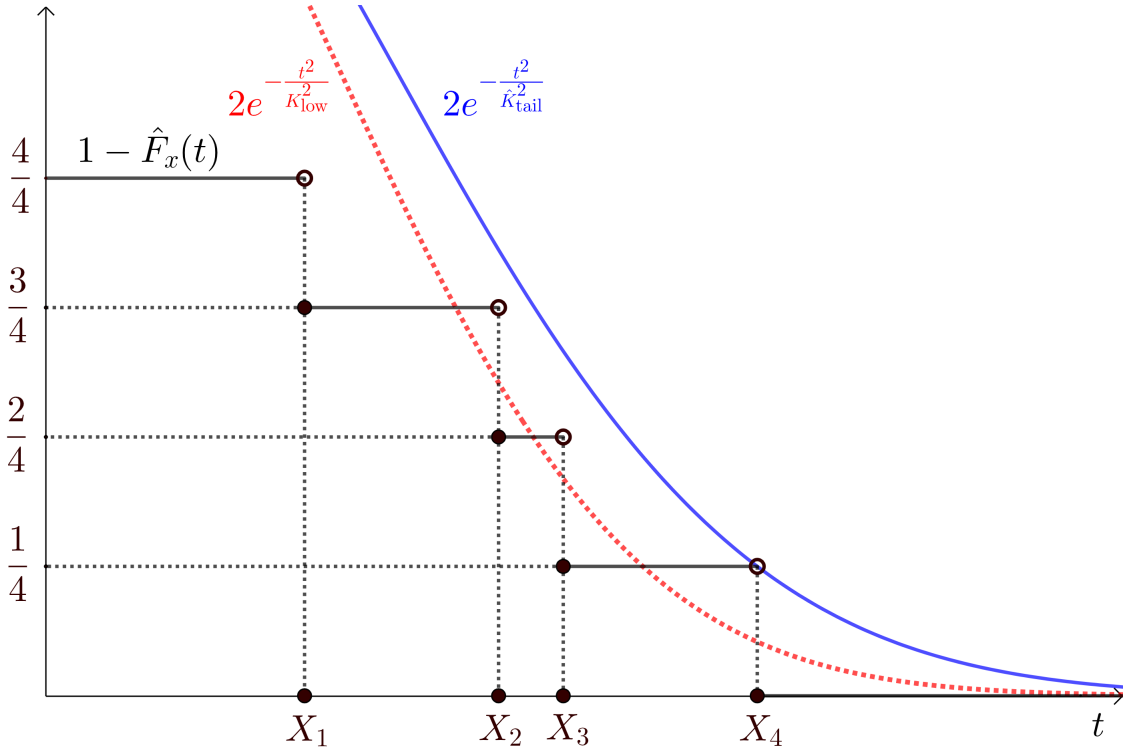
$$\hat{K}_{\text{tail}} = \min_{K \geqslant 0} K \quad \text{s.t.} \quad \forall t \geqslant 0,\ \ 1 - \hat{F}_{|X|}(t) \leqslant 2\exp\left(-\frac{t^2}{K^2}\right) \tag{1.9}$$

$$= \sup_{t \in \mathbb{R}_+} \frac{t}{\sqrt{\log\left(\frac{2}{1-\hat{F}_{|X|}(t)}\right)}} \tag{1.10}$$

$$= \sup_{p \in [0,1]} \frac{\hat{F}^{-1}_{|X|}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}}. \tag{1.11}$$

Program (1.9) has infinitely many constraints so that it appears hard to solve. But the structure of the empirical tail simplifies the solution to finitely many constraints. Consider a plot of an empirical tail for a sample of four (positive) observations X_1, X_2, X_3, X_4 in Figure 1.1. We observe

Figure 1.1: Empirical Tail and Tail Bounds for different parameters



that between realizations X_i and X_{i+1}, the empirical tail is flat. Recall that we are looking for the smallest K such that the empirical tail is uniformly below $t \mapsto 2\exp(-t^2/K^2)$. Since this function is strictly decreasing for any $K \neq 0$, it is sufficient that it is above the empirical tail at $X_1, X_2, \ldots, X_n$. For example, the red dotted line in Figure 1.1 involves a parameter K_{low} that is

too low because the tail bound cuts through the empirical tail at a flat region. In contrast, the blue line is the tail bound for the optimal tail parameter.

With this insight, we can solve for $\hat{K}_{\text{tail}}$ explicitly.

$$\hat{K}_{\text{tail}} = \max_{i=1,\dots,n} \frac{|X_i|}{\sqrt{\log\left(\frac{2}{1-\hat{F}(X_i-)}\right)}}, \tag{1.12}$$

where $\hat{F}(X_i-)$ is defined as the left limit of $\hat{F}$ at X_i. A simpler formula exists when there are no point masses in the distribution of X.[3] Figure 1.1 visualizes the sub-Gaussian tail bound obtained via (1.12). In this case, the maximum in (1.12) is assumed for $i = 4$. This is why the function $t \mapsto 2\exp(-t^2/\hat{K}_{\text{tail}}^2)$ appears to touch the empirical tail at $(X_4, 0.25)$. To be precise, the empirical tail does not pass through that exact point due to its discontinuity at that point.

1.3.3 MGF-sub-Gaussian parameter

Recall that a mean-zero random variable X is MGF-sub-Gaussian with parameter K if for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\exp(\lambda X)] = \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leqslant \exp(K^2\lambda^2).$$

[3] Note that if there are not point masses in the distribution of X, then

$$\hat{K}_{\text{tail}} = \max_{i=1,\dots,n} \frac{|X_{(i)}|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}}, \tag{1.13}$$

where $|X_{(i)}|$ is the i-th order statistic, i.e., the i-th lowest observation among $|X_1|, \dots, |X_n|$.

First note that for $\lambda = 0$, this equality is always satisfied. Rearranging yields that for all $\lambda \in \mathbb{R}$ such that $\lambda \neq 0$

$$\frac{\sqrt{\log\left(\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))]\right)}}{\lambda} \leqslant K.$$

Hence the smallest MGF-sub-Gaussian parameter of X is defined by

$$K^*_{\text{mgf}} := \sup_{\substack{\lambda \in \mathbb{R} \\ \lambda \neq 0}} \frac{\sqrt{\log\left(\mathbb{E}[\exp(\lambda\,(X - \mathbb{E}[X]))]\right)}}{\lambda}. \tag{1.14}$$

This motivates the following estimator for K^*_{mgf}.

$$\hat{K}_{\text{mgf}} := \sup_{\substack{\lambda \in \mathbb{R} \\ \lambda \neq 0}} \frac{\sqrt{\log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(\lambda\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right)\right)}}{\lambda}. \tag{1.15}$$

In contrast to (1.10), there is no obvious closed-form solution for (1.15). So it is relevant to consider its numeric properties. First, evaluation for large (absolute) values of λ involves evaluating the exponential at a large number. To avoid overflow issues, a rescaling can be used. Secondly, it may seem odd to subtract $\mathbb{E}[X]$ in (1.14) since $\mathbb{E}[X] = 0$. However, this is important for the numerical performance of the estimator (1.15) since even when $\mathbb{E}[X] = 0$, it will sometimes be the case that the average is strictly larger than zero. In such a case, not centering X leads to the objective function diverging to infinity in a neighborhood of zero. The issue is resolved by centering. See section 1.9.2 for details.

1.4 Inference

Now that estimators for sub-Gaussian parameters are derived, the natural next question is whether they can be used for inference.

1.4.1 Hoeffding's Inequality

The fundamental theoretical result that I leverage for inference is Hoeffding's inequality. I state it first with known sub-Gaussian parameters as it can be found in the literature.

Theorem 11. Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.3). Fix $n \in \mathbb{N}$. Let $X_1, \ldots, X_n$ be independent, random variables such that for all i, $\mathbb{E}[|X_i|] < \infty$. Assume that for all i, $X_i - \mathbb{E}[X_i]$ is MGF-sub-Gaussian with parameter $K_i > 0$. Then for every $t \geqslant 0$,

$$\mathbb{P}\left[\sqrt{n}\left|\frac{1}{n}\sum_{i=1}(X_i - \mathbb{E}[X_i])\right| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{2^2\frac{1}{n}\sum_{i=1}^{n} K_i^2}\right).^4 \tag{1.16}$$

Since Hoeffding's inequality is a key theoretical result in the context of this paper, it is worth making some observations. First, Hoeffding's inequality is stated with known sub-Gaussian parameters. Extending Hoeffding's inequality to allow for estimated sub-Gaussian parameters is a main objective of this paper.

Second, Hoeffding's inequality is a finite-sample result, i.e., confidence intervals based on (1.16) are valid for any number of observations. In addition, (1.16) is written in such a way that it also allows asymptotic inference.[5] It may not be immediately obvious why asymptotic inference based on Hoeffding's inequality is interesting. After all, inference based on the CLT leads to narrower

[4] More generally, Then for every $t \geqslant 0$, and any $(a_1, \ldots, a_n)$,

$$\mathbb{P}\left[\left|\sum_{i=1} a_i X_i - 0\right| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{2^2\sum_{i=1}^{n} a_i^2 K_i^2}\right). \tag{1.17}$$

. The result in the statement of the theorem then follows from setting $a_i = 1/n$. The more general result 1.17 is sometimes useful in applications.

[5] Asymptotic inference is possible so long as $\frac{1}{n}\sum_{i=1}^{n} K_i^2$ converges.

confidence intervals. The merits of asymptotic inference based on Hoeffding's inequality become clear when the objective is not doing inference on the mean but inference on a function of the mean. Inference based on the CLT requires differentiability in this step as formalized in the delta method. In contrast, inference based on Hoeffding's inequality does not require any smoothness conditions so that it is applicable in situations where CLT-based inference is out of reach. One example of an empirical application in which differentiability is inference for linear programs.

Third, Hoeffding's inequality is robust to heteroskedasticity in the sub-Gaussian parameters.[6] [7] The parameter that governs the tail behavior of the average is the average of the squared sub-Gaussian parameters. It is therefore not necessary to learn all sub-Gaussian parameters individually. It is sufficient to learn about the average of their squares. In the main text, I will assume that the random variables are identically distributed. Appendix section **??** extends this analysis to the case of heteroscedastic sub-Gaussian parameters.

Fourth, Hoeffding's inequality is stated for MGF-sub-Gaussian random variables. The reason for this is that the MGF is a convenient tool to study the sum of independent random variables. That Hoeffding's inequality is stated for MGF-sub-Gaussians might give the impression that it does not apply to tail-sub-Gaussian random variables. Fortunately, one can pass from tail-sub-Gaussians to MGF-sub-Gaussians at the cost of inflating the sub-Gaussian constant. This has been known in the literature (Vershynin, 2018, Proposition 2.5.2). I contribute to this literature by reducing the cost of passing from tail sub-Gaussian parameters to MGF-sub-Gaussian parameters. This allows for sharper inference based on tail-sub-Gaussian parameters.

Proposition 12. Consider a real-valued random variable X with mean zero.

[6] Even if the K_i were equal, the distributions of X_i could differ. This is because sub-Gaussianity only restricts tail behavior.

[7] If the K_i's differ but are bounded, we could always inflate the smaller K_is in light of Remark 10. However, this would result in suboptimal inference.

1. If X is tail-sub-Gaussian with parameter $K > 0$, then X MFG-sub-Gaussian with parameter at most $1.135441K$.

2. If X is MGF-sub-Gaussian with parameter $K > 0$, then X is tail-sub-Gaussian with parameter at most $2K$.

The proof builds on a judicious splitting of the series expansion of two exponentials in a finite and an infinite part and using global polynomial optimization tools to control the finite part. Proposition 12 paves a way to use estimates of tail-sub-Gaussian parameters to conduct inference on MGF-sub-Gaussian parameters and vice versa by inflating the estimates with the appropriate factor. Finally, note that the appearance of 2^2 in Hoeffding's inequality is due to the factor of 2 in Corollary 12 that is required to pass from MGF-sub-Gaussianity to tail-sub-Gaussianity.

Let me conclude this section on Hoeffding's inequality by spelling out the confidence intervals to which it gives rise.

Corollary 13. Fix $n \in \mathbb{N}$. Let $X_1, \dots, X_n$ be independent, random variables such that for all i, $\mathbb{E}[|X_i|] < \infty$. Assume that for all i, $X_i - \mathbb{E}[X_i]$ is MGF-sub-Gaussian with parameter $K_i > 0$. Then for every $\alpha \in (0, 1)$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \in \underbrace{\left[\bar{X}_n - \delta_n(\overline{K^2}, n, \alpha), \bar{X}_n + \delta_n(\overline{K^2}, n, \alpha)\right]}_{=:CI(\bar{X}_n, \overline{K^2}, n, \alpha)}\right] \geqslant 1 - \alpha, \tag{1.18}$$

where

$$\begin{aligned}
\delta_n(\overline{K^2}, n, \alpha) &:= \frac{2}{\sqrt{n}}\sqrt{\overline{K^2}}\sqrt{\log\left(\frac{2}{\alpha}\right)},\\
\bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i,\\
\overline{K^2} &:= \frac{1}{n}\sum_{i=1}^{n} K_i^2.
\end{aligned}$$

The rate at which confidence intervals (CIs) based on Hoeffding's inequality shrink is the optimal rate, $\sqrt{n}$. The difference between CIs based on Hoeffding's inequality and the asymptotically sharp CI based on the CLT is a constant. As discussed in section 1.9.1.1, the 95% CI based of an asymptotically standard normal based on the CLT is $[-1.96, 1.96]$ while the 95% CI of a standard normal based on Hoeffding's inequality is $[-2.72, 2.72]$. This roughly 39% increase in the length of the CI is the price for using Hoeffding's inequality.

1.4.2 Inference with Estimated Parameters

In this section, I study inference based on Hoeffding's inequality with estimated parameters. Consider first inference based on the CLT. When the variance is overestimated, the confidence intervals are larger than they would have to be. In other words, inference is valid but conservative. When the variance is underestimated, the confidence intervals are narrower than they should be. In other words, inference is invalid. The same intuition carries over to inference based on Hoeffding's inequality as the following result shows. The result applies to any estimator of sub-Gaussian parameters.

Theorem 14. Fix $n \in \mathbb{N}$. Let $X_1, \ldots, X_n$ be random variables such that for all i, $\mathbb{E}[|X_i|] < \infty$. Let $\hat{K}$ be any positive random variable with arbitrary dependence on $X_1, \ldots, X_n$.
Then for any $K > 0$ and $\gamma \geqslant 0$

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \in CI(\bar{X}_n, \hat{K}+\gamma, n, \alpha)\right] \\
\geqslant &\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \in CI(\bar{X}_n, K, n, \alpha)\right] + \mathbb{P}\left[\left\{\hat{K}+\gamma \geqslant K^*\right\}\right] - 1,
\end{aligned}$$

where $CI(\cdot,\cdot,\cdot,\cdot)$ is the Hoeffding confidence interval derived in Corollary 13.

1.4.2.1 Asymptotic Inference

The following theorem shows that the probability of underestimating the tail-sub-Gaussian parameter converges to zero as the sample size tends to infinity. Hence, inference based on Hoeffding's inequality with estimated tail-sub-Gaussian parameter is asymptotically valid.

Theorem 15. Suppose that $X_1, \ldots, X_n$ are i.i.d. real-valued tail-sub-Gaussian random variables with minimal parameter K^*. Then for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\hat{K}_{\text{tail}} \leqslant K^*_{\text{tail}} - \varepsilon\right] = 0.$$

In particular, $\hat{K}_{\text{tail}}$ allows for asymptotically conservative inference.

In section **??**, I propose an extension of Theorem 15 to sub-Gaussian parameters that are independent but not identically distributed.

The key step in the proof of Theorem 15 is a localization argument. The estimated tail-sub-Gaussian parameter is the objective value of an estimated function. To guarantee that the maximal value of the estimated function is not far below the maximal value of the population function, it is sufficient that the function is estimated sufficiently well "close" to the population maximizer. Showing that the objective function is sufficiently well estimated "close" to the population maximizer follows from the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality which establishes a finite-sample bound on how close the empirical cdf is to the true cdf.

In contrast, a bound on the probability of overestimating K^* in the generality of Theorem 15 is more challenging. Intuitively, the reason for this is that no obvious localization argument can be applied. The maximum of an estimated function could be overestimated because the estimated function is poorly estimated *anywhere*. While the empirical cdf estimates the true cdf uniformly

well, the empirical quantile function does not achieve uniform consistency to the true quantile function. This uniform consistency is a sufficient condition for asymptotic sharpness of inference with the estimated parameters.

Theorem 16. Suppose that $X_1, \ldots, X_n$ are i.i.d. real-valued tail-sub-Gaussian random variables. Denote the smallest tail parameter with K^*_{tail}. Assume further that the support of X_1 is either finite or bounded and connected. Then for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\hat{K}_{\text{tail}} \geqslant K^*_{\text{tail}} + \varepsilon\right] = 0.$$

In particular, $\hat{K}_{\text{tail}}$ is consistent for K^*_{tail} and inference with $\hat{K}_{\text{tail}}$ is asymptotically sharp in the class of inference methods based on Hoeffding's inequality with estimated tail-sub-Gaussian parameters.

1.4.2.2 Finite-sample Inference

It would be desirable to obtain a finite-sample upper bound on $\mathbb{P}[\hat{K}_{\text{tail}} \geqslant K^*_{\text{tail}}]$ that holds uniformly over all possible tail sub-Gaussian distributions of X with optimal parameter K^*_{tail}. Such a bound would allow finite-sample inference on the sample average with Hoeffding's inequality (1.16) on the basis of Theorem 14. There is no such uniform finite-sample guarantee for the estimator (1.12). In fact, there exists no "reasonable" estimator of K^*_{tail} that satisfies such a uniform finite-sample guarantee as the following theorem shows.

Theorem 17. Consider n i.i.d realizations of a real-valued random variable $X \sim \mu$, $X_1, \ldots, X_n$. Consider any function $\hat{K} = \hat{K}(X_1, \ldots, X_n)$ such that

$$\hat{K}(0, \ldots, 0) = 0. \tag{1.19}$$

Then there exists a μ that is tail-sub-Gaussian with optimal parameter K^*_{tail} such that for any $\varepsilon > 0$

and $0 < \delta < K^*$

$$\mathbb{P}\left[\hat{K} \geqslant K^*_{\text{tail}} - \delta\right] \leqslant \varepsilon.$$

Let me discuss the key condition (1.19). Recall that, the true tail-sub-Gaussian parameter K^*_{tail} of a sub-Gaussian is similar to the standard deviation of a Gaussian. Intuitively, what should be a reasonable estimator of the standard deviation when all observations are zero? More formally, recall that when X is K-tail-sub-Gaussian (or K-MGF-sub-Gaussian), γX is γK-tail-sub-Gaussian (or γK-MGF-sub-Gaussian). Hence, an estimator of K that has a claim to being sharp has to be homogenous of degree one. Homogeneity of degree one implies (1.19). If (1.19) is not satisfied, then $\hat{K}(0, \dots, 0) =: \bar{K} > 0$. In this case, we still get the impossibility result for all $K^*_{\text{tail}} > \bar{K}$.

The proof illustrates the origin of the impossibility result: The uniformity condition includes rather odd distributions with a lot of probability mass at or close to zero and a small amount of probability mass very far away from zero.

To obtain a finite-sample coverage result, we need to rule out such behavior. The key assumption we have to add is an *anti*-concentration assumption. By characterizing the probability of overestimating K^*_{tail} for continuous distributions, the following theorem shows how an anti-concentration can be used for finite-sample inference with estimated concentration parameters.

Theorem 18. Suppose that $X_1, \dots, X_n$ are i.i.d. real-valued tail-sub-Gaussian random variables with parameter K^*. Denote the cdf of X_1 by F and assume that it admits a density f. Consider the estimator $\hat{K}$ defined in (1.13). Then for any $\kappa \geqslant 0$

$$\mathbb{P}\left[\hat{K}_{\text{tail}} \geqslant K^*_{\text{tail}}\kappa\right] = 1 - n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!} F^i(K^*_{\text{tail}} w_{n-i+1,n}(\kappa)) \mathfrak{I}_{n-i,n}, \tag{1.20}$$

where $\mathfrak{I}_{0,n} = 1$, $\mathfrak{I}_{1,n} = F(w_{1,1}(\kappa))$, for $2 \leqslant i \in \mathbb{N}$

$$\mathfrak{I}_{i,n} := \int_0^{w_{1,n}} \dots \int_0^{w_{i,n}} \mathbf{1}_{t_1<\dots<t_i} f(t_1) \dots f(t_i) \mathrm{d}t_i \dots \mathrm{d}t_1,$$

and for any $j, n \in \mathbb{N}$ such that $1 \leqslant j \leqslant n$

$$w_{j,n}(\kappa) := \kappa \sqrt{\log\left(\frac{2}{1 - \frac{j-1}{n}}\right)}.$$

It is remarkable that the continuity assumption allows for a characterization of the probability that K^*_{tail} is overestimated, i.e., that (1.20) is not an inequality. Theorem 18 allows finite-sample inference on $\hat{K}_{\text{tail}}$ if one is willing to add an anti-concentration assumption. Let me illustrate this in a simple example.

Example 19. Consider $n = 2$. Then

$$\begin{aligned}\mathfrak{I}_{0,2} &= 1,\\ \mathfrak{I}_{1,2} &= F(K^*_{\text{tail}} w_{1,2}(\kappa)),\end{aligned}$$

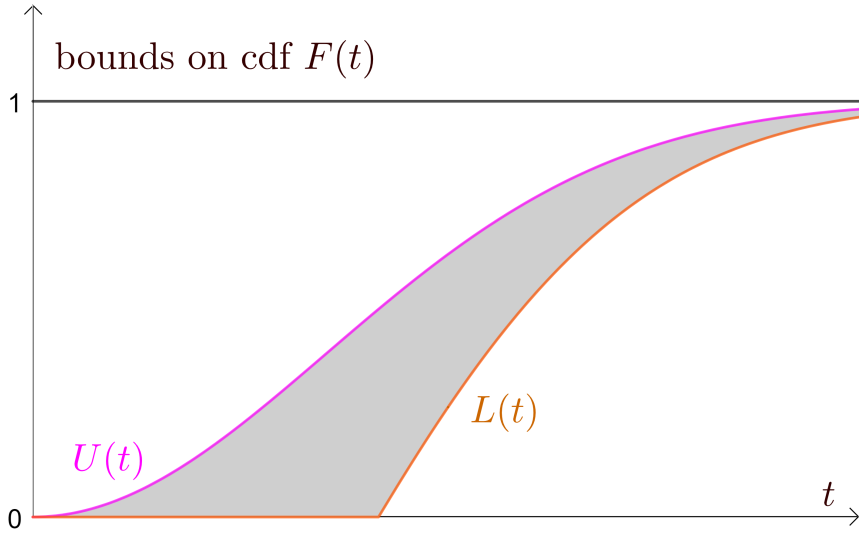
Assume that

$$0 < L\left(\frac{t}{K^*}\right) \leqslant F(t) < U\left(\frac{t}{K^*}\right). \tag{1.21}$$

The sub-Gaussian tail assumption amounts to setting $L(s) = 1 - 2\exp(-s^2)$. If we add an anti-concentration assumption, this gives an upper bound for F. For the purpose of this example, consider $U(s) := 1 - \exp(-s^2)$. Figure 1.2 illustrates this situation. The cdf F must be in the shaded area between the lower bound L and the upper bound U. We see that F has much leeway close to zero but that the tail behavior is tightly bounded from both sides. The tight control is due

to the choice of the bounds for this illustration.

Figure 1.2: Concentration and Anti-Concentration



Using these bounds immediately yields

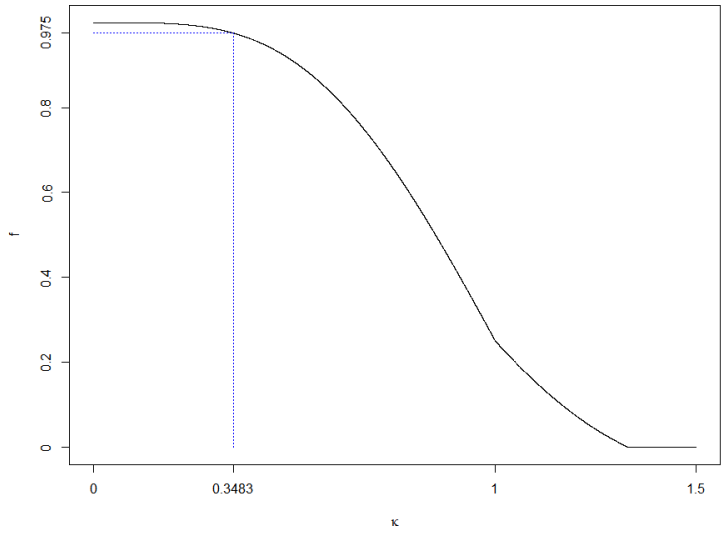
$$\begin{aligned}\mathfrak{I}_{0,2} &\in [1,1]\\ \mathfrak{I}_{1,2} &\in [L(w_{1,2}(\kappa)), U(w_{1,2}(\kappa))]\,,\end{aligned}$$

This allows us to bound (1.20) from below by noting

$$\begin{aligned}&\mathbb{P}\left[\hat{K}_{\text{tail}} \geqslant K^*_{\text{tail}}\kappa\right]\\ =&1 - 2!\sum_{i=1}^{2}\frac{(-1)^{i+1}}{i!}F^i(K^*_{\text{tail}}w_{2-i+1,2}(\kappa))\mathfrak{I}_{2-i,2}\\ =&1 - 2\frac{(-1)^{1+1}}{1!}F^1(K^*_{\text{tail}}w_{2-1+1,2}(\kappa))\mathfrak{I}_{2-1,2} - 2\frac{(-1)^{2+1}}{2!}F^2(K^*_{\text{tail}}w_{2-2+1,2}(\kappa))\mathfrak{I}_{2-2,2}\\ =&1 - 2F^1(K^*_{\text{tail}}w_{2,2}(\kappa))\mathfrak{I}_{1,2} - 2\frac{-1}{2}F^2(K^*_{\text{tail}}w_{1,2}(\kappa))\mathfrak{I}_{0,2}\\ =&1 - 2F^1(K^*_{\text{tail}}w_{2,2}(\kappa))\mathfrak{I}_{1,2} + F^2(K^*_{\text{tail}}w_{1,2}(\kappa))\mathfrak{I}_{0,2}\\ \geqslant&1 - 2U(w_{2,2}(\kappa))U(w_{1,2}(\kappa)) + L^2(w_{1,2}(\kappa)) =: f(\kappa).\end{aligned}$$

The lower bound $f(\kappa)$ for $\mathbb{P}\left[\hat{K}_{\text{tail}} \geqslant K^*_{\text{tail}}\kappa\right]$ is illustrated in Figure 1.3. In particular, for $\kappa \approx 0.3483$, $f(\kappa) = 0.975$.

Figure 1.3: Example Lower Bound for $\mathbb{P}[\hat{K}_{\text{tail}} \geqslant \kappa K^*_{\text{tail}}]$ as function of κ



A valid finite-sample 95% confidence interval with estimated concentration parameters based on Theorem 18 and Theorem 14 would be

$$CI\left(\bar{X}_n, \frac{1}{0.3483}\hat{K}_{\text{tail}}, 2, 0.975\right)$$

since the probability of coverage is bounded from below by $0.975 + 0.975 - 1 = 0.95$. That only the factor $1/0.3483 \approx 2.9$ is sufficient to control for uncertainty in the estimated concentration parameter even though there are only 2 observations is due to the strength of the anti-concentration assumption in this example. □

1.5 Revisiting Applications

1.5.1 Multi-armed Bandits

Consider the multi-armed bandit problem introduced in Example 1. The purpose of this section is to explain the upper confidence bound (UCB) algorithm with known sub-Guassian concentration parameters, and how it can be refined using the estimators derived in section 1.3.

Algorithm 1 UCB with known sub-Gaussian Parameters (Lattimore and Szepesvári, 2020)

1: **Input**: $p_1, \dots, p_L$ (price grid) and $K_{\text{mgf},1}, \dots, K_{\text{mgf},L}$ (MGF-sub-Gaussian parameters for profits)

2: Choose each price **once**: $p_t = p_{\text{grid},t}$ for $t = 1, \dots, L$.

3: Set counters for how often each price was posted $T_1 = 1, \dots, T_L = 1$.

4: Compute the average profit for each price, $\hat{\Pi}_1, \dots, \hat{\Pi}_L$.

5: In period t, choose $p_t = p_{\text{grid},l}$ for any l

$$\underset{l \in \{1,\dots,L\}}{\arg\max} \left(\hat{\Pi}_l + K_l \sqrt{\frac{2 \log\left(1 + t \log^2(t)\right)}{T_l}} \right).$$

6: Increment counter of chosen arm: if $T_l = T_l + 1$.

7: **if** $t < T$ **then**

8: Increment period counter $t = t + 1$.

9: Return to line 4.

Let me begin by recalling the UCB algorithm. The assumption is that for each price $p \in \{p_1, \dots, p_L\}$, the associated profit is random and tail-sub-Gaussian with known tail parameter K_l. After trying each arm once, we can use Hoeffding's inequality to derive confidence intervals for the expected profit for each price of level $\alpha \in (0, 1)$. From then on, the price with the highest upper confidence bound is chosen. Whenever the demand (and hence profit) associated with a price p is observed,

the confidence interval for the profit associated with p is updated. When α is chosen judiciously, it can be shown that this algorithm is asymptotically optimal (Lattimore and Szepesvári, 2020, Theorem 8.1). The asymptotic optimality holds for the optimal sub-Gaussian parameters as well as for inflated parameters, see Remark 10. As market shares are bounded by 1 from above, the support for the profits associated with each price is bounded, allowing to derive possibly inflated sub-Gaussian concentration parameters with Remark 8.

Alternatively, I can use the estimators derived in section 1.3 to estimate the sub-Gaussian concentration parameters "on the fly". In the empirical application presented in section 1.6, I use equation (1.10) to estimate the sub-Gaussian parameter while experimenting. This gives rise to the following slightly different algorithm.

Algorithm 2 UCB with Estimated sub-Gaussian Parameters: Asymptotic Estimators

1: **Input**: $p_1, \ldots, p_L$

2: Choose each price **twice**: $p_t = p_{\text{grid},t-s(t,L)L}$ for $t = 1, \ldots, 2L$ and $s(t, L) = \mathbf{1}_{t>L}$

3: Set counters for how often each price was posted $T_1 = 2, \ldots, T_L = 2$.

4: Compute the average profit for each price, $\hat{\Pi}_1, \ldots, \hat{\Pi}_L$.

5: Estimate the MGF-sub-Gaussian parameters for the profit of each price $\hat{K}_{\text{mgf},1}, \ldots, \hat{K}_{\text{mgf},L}$ with ((1.12) and Proposition 12) or (1.15).

6: In period t, choose $p_t = p_{\text{grid},l}$ for any l

$$\underset{l \in \{1,\ldots,L\}}{\arg\max} \left(\hat{\Pi}_l + \hat{K}_{\text{mgf},1} \sqrt{\frac{2 \log\left(1 + t \log^2(t)\right)}{T_l}} \right).$$

7: Increment counter of chosen arm: if $T_l = T_l + 1$.

8: **if** $t < T$ **then**

9: Increment period counter $t = t + 1$.

10: Return to line 4.

Let me highlight the differences between Algorithm 1 and 2. In contrast to Algorithm 1, Algorithm 2 does not require the sub-Gaussian concentration parameters as inputs. One cost of not knowing sub-Gaussian concentration parameters is that each option has to be tried twice rather than once as there has to be an estimate of the location and of the dispersion of the profit distribution associated with each price.

Note however that theoretical analysis of Algorithm 2 is difficult because finite-sample inference with estimated concentration parameters is impossible in general, see Theorem 17. Under an additional anti-concentration assumption, finite-sample inference with estimated parameters is possible. As a result, a finite-sample regret bound for an appropriately defined variant of the UCB algorithm can be established.

Algorithm 3 UCB with Estimated sub-Gaussian Parameters: Finite-Sample

1: **Input**: $T, p_1, \ldots, p_L, f : \mathbb{N} \to \mathbb{N}$

2: Choose each price **twice**: $p_t = p_{\text{grid},t-s(t,L)L}$ for $t = 1, \ldots, 2L$ and $s(t, L) = \mathbf{1}_{t>L}$

3: Set counters for how often each price was posted $T_1 = 2, \ldots, T_L = 2$.

4: Compute the average profit for each price, $\hat{\Pi}_1, \ldots, \hat{\Pi}_L$.

5: Estimate the MGF-sub-Gaussian parameters for the profit of each price $\hat{K}_{\text{mgf},1}, \ldots, \hat{K}_{\text{mgf},L}$ with (1.10) and Proposition 12.

6: Use Theorem 18 and a union bound to find a κ such that

$$\mathbb{P}[\exists l = 1, \ldots, L \;\; \hat{K}_{\text{mgf},l} < \kappa K_{\text{mgf},l}] \leqslant f(T)\frac{1}{T-L-1}. \tag{1.22}$$

7: In period t, choose $p_t = p_{\text{grid},l}$ for any l

$$\underset{l\in\{1,\ldots,L\}}{\arg\max} \left(\hat{\Pi}_l + \frac{1}{\kappa}\hat{K}_{\text{mgf},l}\sqrt{\frac{2\log\left(1 + t\log^2(t)\right)}{T_l}} \right).$$

8: Increment counter of chosen arm: if $T_l = T_l + 1$.

9: **if** $t < T$ **then**

10: Update the average profit for the chosen arm $\hat{\Pi}_l$.

11: Update the tail-sub-Gaussian parameter $\hat{K}_{\text{mgf},l}$ of the chosen arm l by using Theorem 18 to find a κ such that

$$\mathbb{P}[\hat{K}_{\text{mgf},l} < \kappa K_{\text{mgf},l}] \leqslant f(T)\frac{1}{T-L-1} \tag{1.23}$$

12: Increment period counter $t = t + 1$.

13: Return to line 4.

The main difference between Algorithm 3 and Algorithm 2 is that Algorithm 3 uses the finite-

sample correction factor $\frac{1}{\kappa}$ obtained from Theorem 18 to ensure finite-sample validity of inference based on Hoeffding's inequality with estimated sub-Gaussian parameters. A more subtle difference is that Algorithm 3 takes the number of time periods as input. This means that the firm must know in advance how long it plans to experiment. This is a somewhat undesirable property but it can be relaxed by selecting a period after which there will be no more updating of tail-sub-Gaussian parameters. The key for Algorithm 3 are equation (1.22) and (1.23) as they guarantee that the probability of underestimating a tail-sub-Gaussian parameter is small enough to obtain the following finite-sample regret bound.

Theorem 20. Consider Algorithm 3 in the class of sub-Gaussian bandits with an anti-concentration property that allows the application of Theorem 18 to derive a $\kappa > 0$ that satisfies (1.22) and (1.23). Define the regret of Algorithm 3 by

$$R_T := T\mathbb{E}\left[\Pi^*\right] - \sum_{t=1}^{T} \mathbb{E}\left[\Pi_{l(t)}\right],$$

where $\mathbb{E}\left[\Pi^*\right]$ is the maximum expected profit given the price grid and $l(t)$ is the price chosen by Algorithm 3 in period t. Define the optimality gaps as

$$\Delta_l := \mathbb{E}\left[\Pi^*\right] - \mathbb{E}\left[\Pi_l\right] \geqslant 0.$$

Then

$$R_T \leqslant \sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L} \inf_{\varepsilon\in(0,\Delta_l)} \Delta_l \left(1 + \frac{5}{\varepsilon^2} + \frac{2\left(\log(1+T\log^2(T)+1)\right)}{(\Delta_l-\varepsilon)^2}\right) + f(T)T \max_{l=1,\dots,L} \Delta_l. \quad (1.24)$$

In particular, if $f(T) = 1/T^2$, then the second summand simplifies to $\max_{l=1,\dots,L} \Delta_l/T$.

Let me first discuss what it means for an anti-concentration property to allow the application of Theorem 18 to derive a $\kappa > 0$ that satisfies (1.22) and (1.23). Theorem 18 allows to bound

the probability of underestimating the tail-sub-Gaussian parameter in finite samples as long as the distribution for which one estimates the tail-sub-Gaussian parameter satisfies an anti-concentration assumption. The anti-concentration condition in (1.21) in Example 19 illustrates this. The stronger the anti-concentration, the less do I have to inflate the estimated sub-Gaussian tail parameters by multiplying them with $\frac{1}{\kappa}$. Without an anti-concentration property, Theorem 18 shows that it is impossible to conduct finite-sample inference with estimated sub-Gaussian tail parameters.

Next, note that choosing $f(T)$ in Algorithm 3 involves a trade-off: choosing a smaller f reduces the second summand in (1.24) but increases the first summand in (1.24) as reducing the probability of underestimating a tail-sub-Gaussian parameter comes at the cost of a higher correction factor $\frac{1}{\kappa}$.

Finally, note that (1.24) is a finite-sample bound that gets close to the finite-sample regret bound of Theorem 8.1 in (Lattimore and Szepesvári, 2020). Lattimore and Szepesvári (2020) use the finite-sample regret bound to show that the regret of Algorithm 1 reaches an asymptotic lower bound in Theorem 15.2 in (Lattimore and Szepesvári, 2020) and can hence be considered asymptotically optimal in the minimax-sense. The difference between the regret bound (1.24) and the regret bound in Theorem 8.1 of (Lattimore and Szepesvári, 2020) is the second summand that can be made arbitrarily small. To derive asymptotic regret bounds, one has to adapt Algorithm 3 by choosing a period after which the sub-Gaussian parameters are not further updated.

Let me complement these theoretical results with a simulation study. I focus on Algorithm 1 and 2. For illustrative purposes, I suppose that there are only 2 prices, p_1 and p_2. I further assume that the profit associated with price p_1 is distributed according to the $U[0,1]$ distribution. For some $\Delta > 0$, I assume that the profit associated with price p_2 is distributed according to the $U[\Delta, 1+\Delta]$ distribution. In particular, p_2 is maximizing expected profits. The higher Δ, the more does price p_2 increase expected profits. However, note that both profit distributions coincide when centered

so that they share the same optimal tail-sub-Gaussian parameter K^*_{tail}. For notational simplicity, denote $K^* := K^*_{\text{tail}}$ and $\hat{K}_{\text{tail},l} = \hat{K}_L$. I consider a learning horizon of $T = 100$ periods. A first

Table 1.1: Influence of Concentration Parameters K_1, K_2 on Profits

	UCB1					UCB2
Δ	$K_1 = K^*$ $K_2 = K^*$	$K_1 = 10K^*$ $K_2 = 10K^*$	$K_1 = 0.1K^*$ $K_2 = 0.1K^*$	$K_1 = K^*$ $K_2 = 10K^*$	$K_1 = K^*$ $K_2 = 0.1K^*$	$K_i = \hat{K}_1$ $K_2 = \hat{K}_2$
0.05	3.069 (0.01)	2.564 (0.00)	3.084 (0.02)	4.928 (0.00)	0.254 (0.00)	2.873 (0.01)
0.10	7.123 (0.01)	5.253 (0.00)	7.134 (0.04)	9.861 (0.00)	0.861 (0.01)	6.429 (0.03)
0.15	11.777 (0.01)	8.074 (0.00)	11.920 (0.05)	14.800 (0.00)	2.325 (0.03)	10.531 (0.04)
0.20	16.709 (0.01)	11.017 (0.00)	17.158 (0.06)	19.743 (0.00)	5.097 (0.06)	15.023 (0.05)
0.25	21.793 (0.01)	14.082 (0.00)	22.660 (0.06)	24.691 (0.00)	9.296 (0.08)	19.832 (0.06)
0.30	26.876 (0.01)	17.264 (0.00)	28.209 (0.05)	29.645 (0.00)	14.559 (0.10)	24.851 (0.06)

Notes: The table reports the the sum of expected profits over $T = 100$ periods when using various UCB algorithms to experiment with prices. The first column is the mean difference in expected profits between the two prices. Column 2-5 are the expected profits of UCB Algorithm 1. UCB Algorithm 1 requires the specification of sub-Gaussian parameters K_1, K_2. Column 2 reports expected profits of the UCB algorithm with the optimal concentration parameters, column 3 reports expected profits of the UCB algorithm with too high concentration parameters, column 4 reports expected profits of the UCB algorithm with too low concentration parameters. Column 4 looks at the asymmetric case when the concentration parameter of the optimal second arm is inflated while the concentration parameter of the sub-optimal first arm is correct. Column 4 looks at the asymmetric case when the concentration parameter of the optimal second arm is too low while the concentration parameter of the sub-optimal first arm is correct. Each number is based on 10000 simulations. The numbers in brackets below the expected profits are simulation standard deviations.

observation is that the expected sum of profits attained by UCB Algorithm 1 with the smallest concentration parameters in column 2 are higher than the expected sum of profits attained by UCB Algorithm 1 with inflated concentration parameters in column 3. The reason for this is that inflated sub-Gaussian parameters give too much room for optimism, which leads to too much exploration

by the UCB algorithm. This observation illustrates a finite-sample issue of the regret bound in Theorem 8.1 in Lattimore and Szepesvári (2020): it shows asymptotic optimality of the UCB algorithm as long as the true concentration parameters are not underestimated. The simulation highlights the importance of being close to the smallest possible concentration parameters in finite samples.

A second observation is that for all sub-optimality gaps Δ, UCB 1 with $K_2 = 10K^*$ and $K_1 = K^*$ leads to the highest profits. This is not surprising once we recall that the UCB algorithm recommends the option when it has a high mean or a high uncertainty. When the measure of uncertainty in the rewards associated with the optimal parameter is large, this gives rise to a strong motive for exploring this option - almost regardless of the mean. Of course, it is not feasible to inflate the concentration parameter for the profits of the profit-maximing price since the identity of the profit-maximing price is unknown. The flip-side of this phenomenon is when the concentration parameter of the optimal arm is too low. This leads to an overexploration of the sub-optimal price because there is too much optimism about it.

A third observation is that the expected sum of profits achieved by UCB Algorithm 2 which estimates sub-Gaussian parameters is close to the expected sum of profits achieved by UCB Algorithm 1 with optimal concentration parameters.

1.5.2 Linear Programs

Consider the linear program

$$\min_{x \in \mathbb{R}^p} c'x \qquad \text{s.t.} \qquad Ax \leqslant b \tag{1.25}$$

with constraint matrix $A \in \mathbb{R}^{m \times p}$, constraint vector $b \in \mathbb{R}^m$ and objective vector $c \in \mathbb{R}^p$. The objective of this section is to derive methods for inference on the optimal value $c'x^*$ of this linear

program when A, b and c are expectations that have to be estimated. It is instructive to start by considering the case where only b is estimated and then successively extend the analysis to the case where c and A are also estimated.

Let me first consider inference on the objective value of (1.25).

Theorem 21. Consider the linear program (1.25). Assume that for all $j = 1, \dots, p$ and and for all $k = 1, \dots, m$

$$\begin{aligned} A_{k,j} &= \mathbb{E}[\check{A}_{k,j}], \\ b_k &= \mathbb{E}[\check{b}_k], \\ c_j &= \mathbb{E}[\check{b}_j], \end{aligned}$$

where $\check{A}_{k,j}$, $\check{b}_k$, and $\check{c}_j$ are real-valued random variables, which are MGF-sub-Gaussian with parameter $K_{A,k,j}$, $K_{b,k}$, and $K_{c,j}$. Suppose further that there is an i.i.d. sample $(\check{A}^i, \check{b}^i, \check{c}^i)_{i=1,\dots,n}$ of $(\check{A}, \check{b}, \check{c})$. Denote for any k, j

$$\begin{aligned} \bar{A}_{k,j} &= \frac{1}{n} \sum_{i=1}^{n} \check{A}^i_{k,j}, \\ \bar{b}_k &= \frac{1}{n} \sum_{i=1}^{n} \check{b}^i_k, \\ \bar{c}_j &= \frac{1}{n} \sum_{i=1}^{n} \check{b}^i_j. \end{aligned}$$

Fix $\alpha \in (0, 1)$. There are $(\delta_{A,k,j,n,K_{A,k,j}}, \delta_{b,k,n,K_{b,k}}, \delta_{c,j,n,K_{c,j}})_{k \in \{1,\dots,m\}, j \in \{1,\dots,p\}}$ such that

$$\begin{aligned} 1 - \alpha \leqslant \mathbb{P}[\forall j, k \quad & A_{k,j} \in [\bar{A}_{k,j} - \delta_{A,k,j,n,K_{A,k,j}}, \bar{A}_{k,j} + \delta_{A,k,j,n,K_{A,k,j}}], \\ & b_k \in [\bar{b}_k - \delta_{b,k,n,K_{b,k}}, \bar{b}_k + \delta_{b,k,n,K_{b,k}}], \\ & c_j \in [\bar{c}_j - \delta_{c,j,n,K_{c,j}}, \bar{c}_j + \delta_{c,j,n,K_{c,j}}]]. \end{aligned}$$

Then the quadratic program with quadratic constraints

$$\min_{\substack{x\in\mathbb{R}^p,\\ \tilde{A}\in\mathbb{R}^{m\times p},\\ \tilde{b}\in\mathbb{R}^m,\\ \tilde{c}\in\mathbb{R}^p}} \frac{1}{2}(\tilde{c}', x') \underbrace{\begin{pmatrix} 0_{p\times p} & I_p \\ I_p & 0_{p\times p} \end{pmatrix}}_{=:M} \begin{pmatrix} \tilde{c} \\ x \end{pmatrix} \text{ s.t. } \begin{cases} \tilde{A}x \leqslant \tilde{b}, \\ \tilde{A}_{k,j} \in [\bar{A}_{k,j} - \delta_{A,k,j,n,K_{A,k,j}}, \bar{A}_{k,j} + \delta_{A,k,j,n,K_{A,k,j}}], \\ \tilde{b}_k \in [\bar{b}_k - \delta_{b,k,n,K_{b,k}}, \bar{b}_k + \delta_{b,k,n,K_{b,k}}], \\ \tilde{c}_j \in [\bar{c}_j - \delta_{c,j,n,K_{c,j}}, \bar{c}_j + \delta_{c,j,n,K_{c,j}}] \end{cases} \tag{1.26}$$

yields a lower bound for the α quantile of the objective value of (1.25).

Theorem 21 establishes the validity of inference on the optimal value based on Hoeffding's inequality. In principle, any finite-sample inference method can be used to achieve the same guarantee. The advantage of Hoeffding's inequality is that (1.26) is computationally attractive in many important examples. Before I discuss the computational properties of (1.26) in detail, let me first note that there are many cases where only a subset of parameters is estimated. For example, Fang et al. (2022) consider the case where only b is estimated while A and c are known and list numerous economic applications. If, for instance, parameter $A_{k,j}$ is known for some k, j, one can simply set $\delta_{A,k,j,n,K_{A,k,j}} = 0$.

The following result discusses the computational complexity of (1.26).

Theorem 22. Consider the quadratically constrained quadratic program (1.26).

1. If only b is estimated, (1.26) reduces to a linear program.

2. If b and c are estimated but A is known, (1.26) reduces to a quadratic program. In general, (1.26) is strongly NP-hard. A sufficient condition for convexity is that x and c are non-negative (or that both are non-positive).

3. If A, b and c are estimated, (1.26) is strongly NP-hard in general. A sufficient condition for convexity is that x and c are non-negative and that A is known to be non-positive.

1.5.2.1 Application to Nonparametric Demand Estimation

One empirical setting to which this can be applied is nonparametric demand estimation. Tebaldi et al. (2019) proposed a method to compute sharp bounds on demand counterfactuals without parametric assumptions. In particular, they derive sharp bounds for the drop in health insurance coverage rates if premium subsidies were lowered. To determine these bounds when prices are exogenous, they solve the following linear program (LP).

$$\min_{\phi} c'\phi \qquad \begin{cases} \phi_k & \geqslant 0, \\ \sum_k \phi_k & = 1, \\ \sum_{k \in S(j,m)} \phi_k & = \hat{s}_{j,m}. \end{cases} \tag{1.27}$$

Here, the ϕ_k are probability masses that an unknown distribution of latent utility draws puts on elements of the Minimum Relevant Partition (MRP). The MRP is a partition of the space of latent utility draws and only depends on the observed prices. Consider a case where there is only one geographical market that is observed over time. Assume that the distribution of latent utility draws does not change over time. Then we observe the choices of the same population over time when it faces different prices. The MRP is the coarsest partition of the space of latent utility draws such that two consumers choose the same (health insurance) option if and only if their latent utility draws lie in the same partition element. The objective c formalizes an objective set by the researcher, e.g., the demand for health insurance when a subsidy is decreased. $S(j, m)$ is a set of indices of MRP elements that make up the observed market shares.

Since the prices are observed without error, everything that depends only on prices (the MRP,

$S(j,m)$) can be constructed without error. However, there may be error in the estimated market shares $\hat{s}_{j,m}$. When there are only finitely many consumers and consumers make independent decisions, errors in the estimated market shares are inevitable. How much does the estimation error in market shares propagate to error in the bounds derived by Tebaldi et al. (2019)?

I explore this in a numerical simulation with the following data generating process. Suppose that there are 2 inside products differentiated only by their price in $T = 10$ or $T = 20$ markets. I assume that consumer i in market t choose the product $j \in \{0, 1, 2\}$ that maximizes utility $u_{i,j,t}$, where

$$u_{i,j,t} = -\alpha p_{j,t} + \varepsilon_{i,j,t}.$$

I further assume that α, the marginal utility of income, is given by $\alpha = 1$, and that the latent utility draws $\varepsilon_{i,j,t}$ follow a T1EV distribution. For prices, I simply draw from the uniform distribution between 0 and 1. The utility of the outside option is normalized to $u_{i,0,t} = 0$.

The market shares are thus described by the usual logit formula, i.e.,

$$s_{j,t} = \frac{\exp(-p_{j,t})}{1 + \exp(-p_{1,t}) + \exp(-p_{2,t})}$$

Denote the number of consumers by $n \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6, \}$. We assume that the *realized* market shares are drawn from a binomial distribution where the number of experiments is the number of consumers n and the success probability is the market share implied by the logit formula. The binomial draws are the only source of randomness in this simulation.

For the objective vector c in (1.27), I simply draw from the uniform distribution on the unit interval. For $T = 10$, this gives rise to a Linear Program with 66 variables, 20 constraints with

estimated parameter and 67 shape constraints that are not influenced by the estimation of market shares. For $T = 20$, the Linear Program has 231 variables, 40 constraints with estimated parameter and 232 shape constraints that are not influenced by the estimation of market shares.

The linear program that is solved for inference is then

$$\min_{\phi} c'\phi \qquad \begin{cases} \phi_k & \geqslant 0, \\ \sum_k \phi_k & = 1, \\ \sum_{k \in S(j,m)} \phi_k & \leqslant \hat{s}_{j,m} - \delta(j, m, K_{j,m}, \alpha) \\ \sum_{k \in S(j,m)} \phi_k & \leqslant \hat{s}_{j,m} + \delta(j, m, K_{j,m}, \alpha), \end{cases} \tag{1.28}$$

where $2\delta(j, m, K_{j,m}, \alpha)$ is the length of the two-sided CI for $\hat{s}_{j,m}$ based on Hoeffding's inequality with MGF-sub-Gaussian parameter $K_{j,m}$ where $K_{j,m}$, see Corollary 13. The dependence on the number of consumers is 'hidden' in the dependence on the market. To estimate $K_{j,m}$ from just the market shares, I create a binary vector with as many 1s as there are sold products (i.e., the product of observed market shares and number of consumers). Then I can use either the estimator for the tail-sub-Gaussian parameter 1.10 and inflate it with Proposition 12, or directly use the MGF-sub-Gaussian parameter 1.15. Both will yield asymptotically valid inference based on Theorem 14 and Theorem 15. From a computational perspective, it is important to note that (1.28) is computationally attractive because δ does not depend on ϕ. This is a feature of Hoeffding's inequality. Other finite-sample methods, such as Chernoff's inequality or binomial quantiles would also allow for valid inference. But the nonlinear dependence of the implied δ would render (1.28) intractable when there are many parameters and inequalities.

The results are reported in Table 1.2. We see that the inference based on Hoeffding's inequality is conservative, coverage rates for all numbers of markets and consumers are 100%, vastly overcov-

ering the required 95%. Still, the bounds are informative in that they exclude a fair share of the unit interval. We also see that estimating concentration parameters leads to narrower confidence intervals compared to the using support bounds with Remark 8. Inference with MGF-sub-Gaussian parameters also is less conservative than inference with tail-sub-Gaussian parameters due to the inflation factor from Proposition 12.

Table 1.2: Inference for Linear Programs: Simulation for Nonparametric Demand Counterfactuals

			quantiles				
				Hoeffding with tail		Hoeffding with MGF	
T	n	c^*	simulated	support bd	estimated	support bd	estimated
10	10^5	72.40%	73.24%	85.43%	79.63%	77.73%	77.51%
			(0.95)	(1.00)	(1.00)	(1.00)	(1.00)
10	10^6	72.61%	72.95%	79.25%	76.79%	75.33%	75.17%
			(0.95)	(1.00)	(1.00)	(1.00)	(1.00)
20	10^5	54.73%	55.89%	81.43%	66.74%	62.63%	62.26%
			(0.95)	(1.00)	(1.00)	(1.00)	(1.00)
20	10^6	55.14%	55.64%	65.81%	61.01%	59.13%	58.93%
			(0.95)	(1.00)	(1.00)	(1.00)	(1.00)

Notes: The table reports a simulation results of (1.27)*. T is the number of markets, n is the number of consumers, c* is the average optimal value across simulations. The other columns then report the 0.95 quantiles of this objective value. The exact quantiles based on simulation are in column 4. The other columns report bounds on the quantiles based on Hoeffding's inequality, with tail-sub-Gaussian parameters based on support bounds in column 5, estimated parameters in column 6; MGF-sub-Gaussian parameters based on support bounds in column 7, estimated parameters in column 8. Below the quantiles, I report simulated coverage rates. Each number is based on 100 simulations.*

In addition to allowing for inference, (1.28) is feasible with probability of at least $1 - \alpha$ as long as the sub-Gaussian parameter is valid. In contrast, Tebaldi et al. (2019) equate the choice shares with the estimated market shares, leading to infeasibility of (1.27). To accommodate potential

infeasibility, 1.3 propose a two-step method. As Table 1.3 shows, feasibility is not a practical concern with (1.28), echoing the discussion of finite-sample feasibilities in other popular demand models (Lieber and Wiemann, 2022). That all instances of (1.3) are feasible could be an artifact of the conservativeness of Hoeffding's inequality or it could be a result of (1.27) being "just" infeasible.

Table 1.3: Nonparametric Bounds on Demand Counterfactuals: Feasibility

			Fraction of Feasible Draws			
			Hoeffding with tail		Hoeffding with MGF	
T	n	$s = \hat{s}$	support bd	estimated	support bd	estimated
10	10^2	0.00	1.00	1.00	1.00	1.00
10	10^3	0.16	1.00	1.00	1.00	1.00
10	10^4	0.95	1.00	1.00	1.00	1.00
10	10^5	1.00	1.00	1.00	1.00	1.00
20	10^2	0.00	1.00	1.00	1.00	1.00
20	10^3	0.00	1.00	1.00	1.00	1.00
20	10^4	0.39	1.00	1.00	1.00	1.00
20	10^5	0.98	1.00	1.00	1.00	1.00

Notes: The table reports the how often the program (1.27) *was feasible in the simulation.T is the number of markets, n is the number of consumers. Column 3 reports how often the linear program* (1.27) *that equates theoretical choice probabilities s and observed market shares $\hat{s}$. Column 4-7 report how often the program* (1.28) *is feasible if Hoeffding's inequality with tail-parameters (based on support bounds or estimation) or MGF-parameters (based on support bounds or estimation) are used. Each number is based on 100 simulations.*

1.6 Empirical Application: Liquor Sales in Washington State

In November 2011, voters in Washington State adopted Initiative 1183, ending a state-monopoly for liquor sales in Washington that had existed for almost 80 years. Starting June 1st, 2012, vendors were free to sell liquor and to set prices if they bought a license from the Washington State Liquor Control Board. To set profit-maximizing prices, firms have to know the demand curve. Anecdotal evidence suggests that vendors did not know the demand curve. Alan Johnson, CEO of BevMo!, a Californian company specializing in the sale of alcoholic beverages that entered the liquor markets in Washington, said in March 2012:

> *"I sure don't know what we'll charge the consumer.*
> *There is going to be a lot of scrambling."*[8]

In this situation, a firm can learn the demand curve by simply setting prices then observing the corresponding demand. When there are only finitely many consumers, the observed market share will only be a approximation of the theoretical market share. In other words, the firm can learn about the demand curve by collecting noisy signals of it. But learning the demand curve is only a means to an end. The firm's objective is to maximize profits. This problem can be modeled as a multi-armed bandit (Rothschild, 1974).

The objective of this section is to quantify how well the entrants into the off-premise liquor sales learned to set prices. Could they have increased profits had they just used the UCB algorithm? If so, by how much? Does it depend on which concentration parameters are supplied to the UCB algorithm?

For this empirical exercise, I proceed in three steps. First, I estimate the demand for liquor. In

[8]Cited according to Huang et al. (2022) as the original source Gregutt, Paul, "BevMo Ramps It Up in Washington State" (April 30, 2012) is no longer available.

a second step, I estimate the marginal cost of each liquor. In a third step, I compare the price experimentation implemented by the store managers to three different UCB algorithms: the UCB algorithm with concentration parameters based on support bound, the UCB algorithm with optimal concentration parameters, and the UCB algorithm with estimated concentration parameters.

1.6.1 Data

I combine data from three sources. My main data source is the Retail Scanner Data from NielsenIQ which provides weekly price, marketing and sales information of liquor from June 2012 to December 2014. I complement this with monthly liquor price data from Oregon which helps to obtain estimates for marginal cost. The 2010 Census contains demographic information the a county level. In particular, it contains the number of residents above the federal minimum legal drinking age of 21 years by county.

1.6.2 Sample Construction

The main liquor categories are whiskey, gin, rum, tequila and vodka. Consumers rarely substitute across product categories so that firms tend to disregard substitution effects across product categories (Conlon and Rao, 2015). As Huang et al. (2022), I focus on whiskeys as it is the category with the highest sales.

There is evidence that consumers in Washington do not travel between stores to shop for liquor (Illanes and Moshary, 2020; Huang et al., 2022). To be sure that consumers do not travel between stores, I restrict my sample to counties in which a liquor stores held a local monopoly. The purpose of this restriction is to circumvent strategic interactions between competing off-premise liquor stores. When there is competition, an increase in demand could stem from noise, one's own experimental reduction of prices or from a competitor's price increase. There are 4 counties in

Washington in which a chain of liquor stores held monopolies.

The NielsenIQ retail scanner data set does not record prices of products that were not sold in a given store in a given week (Dubé et al., 2021). Observing prices is necessary to evaluate store managers' experimentation schemes. Hence, I focus on the products with sufficiently high sales so that I can observe posted prices reliably. Specifically, I focus on the 10 whiskeys with highest sales. Even with this restriction, there are store-week pairs with zero sales. When sales for a whiskey in a store-week are zero, I use the uniform pricing policy within retail chains (DellaVigna and Gentzkow, 2019; Huang et al., 2022) to recover the unobserved prices.

1.6.3 Demand Estimation

I assume that demand follows a standard simple logit. Specifically, I model the utility for whiskey j in market t as

$$u_{j,m,t} = x_{j,t}\beta - p_{j,t}\alpha + \delta_m + \xi_{j,t} + \varepsilon_{j,t},$$

where $x_{j,t}$ contain brand, size, whiskey type, year, month fixed effects, $p_{j,t}$ is the price of product j in market t and δ_m are store fixed effects. The utility of the outside option (a composite between choosing no liquor and choosing another liquor) is normalized to zero. Markets are defined on a county-week level. Then the number of sales of product j in market t follows a binomial distribution where the number of trials is the number of consumers in the market and the success probability is the choice probability

$$s_{j,m,t} = \frac{\exp\left(x_{j,t}\beta - p_{j,t}\alpha + \delta_m + \xi_{j,t}\right)}{\sum_{k=0}^{J} \exp\left(x_{k,t}\beta - p_{k,t}\alpha + \delta_m + \xi_{k,t}\right)}.$$

Because the county-week level is a relatively disaggregated definition of markets, there are many liquors and few consumers. Hence, there are markets with zero market shares. I therefore apply the method developed in Lieber and Wiemann (2022) to estimate the demand model as it is robust to zero-valued market shares. I use a confidence level of 0.95. Prices are allowed to be endogenous. As instruments, I use the prices in Oregon in the same month[9] as well as average prices in Washington in the same week in other markets, not just from the four counties in the sample.

It is worth discussing exactly what the price $p_{j,t}$ is. For tax purposes, Washington State considers a liquor as spirit whenever the alcohol by volume is above 20%. This is the case for all whiskeys in my sample. There are three relevant taxes. First, there is a sales tax of 20.5% on each liquor. Second, there is a tax of \$3.7708 on each liter of spirit sold[10]. Finally, there is a liquor licensing fee of 17% of revenue that liquor stores have to pay to the state. Before the liberalization of the liquor market, all taxes were included in the sticker price on the shelf. After the liberalization, liquor stores initially posted sticker prices that do not include the sales tax of 20.5% or the liter tax, confusing customers. There was a push towards including taxes in sticker prices.[11] It is hard to say which vendor included which tax on the price tag when.[12] But since this is a widely discussed issue in the media at the time, the WSLCB provides clear information on its website[13] and there is even an online calculator[14] to calculate the final price, I assume that consumers are aware of the full final price.

Whiskey is a durable good. It is unlikely to depreciate in quality within a week when properly

[9]There are months in which I do not observe prices in Oregon. I use locally constant and linear interpolation to interpolate the prices in months for which prices in Oregon are not observed.

[10]This applies to the volume of spirits, not to the volume of alcohol.

[11]Melissa Allison, " Retailers simplifying liquor-price tags to include taxes ", Seattle Times, July 5, 2012. https://www.seattletimes.com/business/retailers-simplifying-liquor-price-tags-to-include-taxes/

[12]I have ordered liquor "for pick up" in Seattle myself. No online store included the taxes in the sticker price.

[13]See, e.g., https://dor.wa.gov/taxes-rates/other-taxes/spirits-hard-liquor-sales-tax or https://dor.wa.gov/about/statistics-reports/spirits-taxes , last accessed 10/03/2022.

[14]https://www.liquorcalc.com/ , last accessed 10/03/2022

stored so that a whiskey bottle purchased in one week might be consumed in a later week. While it would be possible to incorporate this durability for demand estimation by considering a dynamic program in which the state variable is the volume (and type) of whiskey stored by a consumer, this creates a non-stationary learning environment for the firm engaging in price experimentation. I therefore abstract from the durability of whiskey in the demand estimation.

Table 1.4: Demand Estimates

	main specification	price (first stage)
price	-0.089	
feature or display	0.146	-0.366 (0.693)
missing feature & display	1.342	-0.411 (0.139)
bourbon straight bonded	0.658	-5.507 (0.755)
extra large (1.75l vs 0.75l)	1.240	0.855 (0.474)
intercept	-7.786	-1.252 (0.523)
year fixed effects	✓	✓
month fixed effects	✓	✓
brand fixed effects	✓	✓
store fixed effects	✓	✓
avg. prices in Washington		1.549 (0.056)
prices in Oregon		0.070 (0.020)
number of Observations	5360	5360

Notes: The first column provides logit parameter estimates following the methodology in Lieber and Wiemann (2022). The second column reports estimates of the first stage for price. The F-statistic for the two excluded instruments is 391.49.

The parameter estimates are reported in Table 1.4. The implied average price elasticities for 6 of the 10 whiskeys are reported in Table 1.5. The pattern of the cross-price elasticity exhibits the Independence-of-Irrelevant-Alternatives property of the simple logit. Cross-price elasticities are very low. In contrast, own-price elasticities are considerable and suggest that prices are in the elastic range of the demand curve. I have chosen to report the elasticities for only 6 of the 10 whiskeys for formatting reasons, because I cannot identify the products and because the pattern

for the 10 products is the same. I can compare these elasticities to those reported by Huang et al. (2022) who estimate a random coefficient logit model with the same data but more products (sales aggregated to a monthly level). The own-price elasticities are of comparable size, while the cross-price elasticities reported by Huang et al. (2022) are higher by a factor of around 10. Even with those cross-price elasticities, there seems to be limited substitution for marginal price changes.

Table 1.5: Average Price Elasticities

	Elasticity with respect to price of product					
	1	2	3	4	5	6
product						
1	-2.3006	0.0007	0.0008	0.0007	0.0016	0.0006
2	0.0006	-2.4477	0.0008	0.0007	0.0016	0.0006
3	0.0006	0.0007	-3.1063	0.0007	0.0016	0.0006
4	0.0006	0.0007	0.0008	-3.0305	0.0016	0.0006
5	0.0006	0.0007	0.0008	0.0007	-2.2121	0.0006
6	0.0006	0.0007	0.0008	0.0007	0.0016	-1.1499

Notes: The table reports the average price elasticities for 6 whiskeys. To interpret these estimates, consider the first row. It is the elasticity of the market share of whiskey 1 with respect to the prices of whiskeys 1-6 when all product characteristics and prices are set to their average across regions and weeks from June 2012 to December 2014. Intuitively, a 1% increase in the price of product 1 will reduce its market share by 2.3% and increase the market share of whiskey 2 by 0.0007%.

1.6.4 Marginal Cost

To obtain these estimates for marginal cost, one would usually assume that firms play a Nash-Bertrand equilibrium. Then one could derive the first order conditions and back out the marginal

cost parameters. However, if firms experiment with prices, marginal costs cannot be obtained by an inversion of first order conditions.

An institutional detail helps to overcome this problem: Oregon, a neighbor of Washington State, maintains a state monopoly on off-premise liquor sales and has regulated prices to be the wholesale prices plus a fixed 104% markup. Since prices in both Oregon and Washington State include federal excise taxes, there are no tax differences in the acquisition of liquor. This allows me to back out the wholesale prices in Oregon (including federal excise tax), which are then assumed to equal wholesale prices (including federal excise tax) in Washington State.

1.6.5 Model of Firm Experimentation Behavior

Each monopoly's objective is to maximize the sum of expected profits[15] over the period from June 2012 to December 2014 by choosing a sequence of prices $(p_{j,t})_{j\in\{1,\dots,J\},t\in\{1,\dots,T\}}$. $p_{j,m,t}$ is allowed to depend on the information available to monopoly m up to period $t-1$.

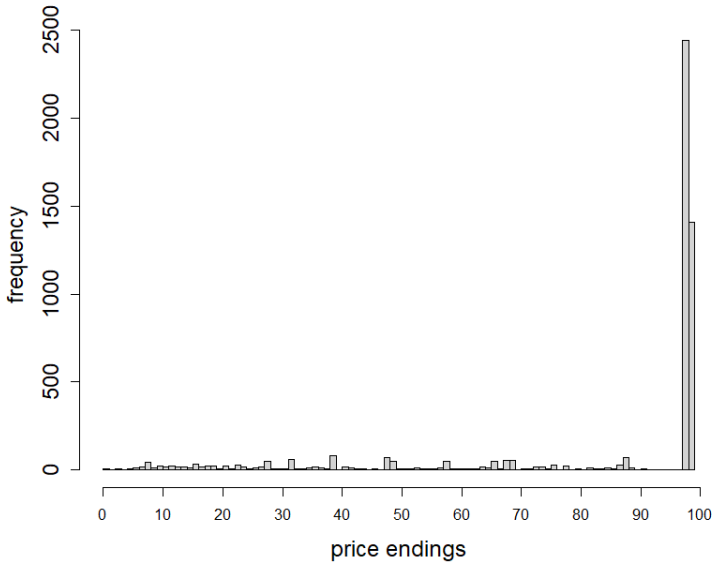
I assume that monopolies only observe the realizations of demand in their own markets, i.e., that they do not observe the demand realizations of other monopolists. This would incentivize free-riding on other monopolies' experimentation efforts. Whether the group of monopolies would then experiment enough is a very interesting question (Che and Hörner, 2018).

I define the length of a period to be one week, since that is the level of temporary granularity at which I observe prices in the retail scanner data. It might be that firms change their prices within a week. NielsenIQ then reports a quantity weighted average of prices. I observe some patterns in

[15] In the bandit literature, it is common to minimize the *regret*, i.e., difference the expected profit accruing from charging the optimal price in every period and the expected profit of the actually implemented price experimentation strategy. Benefits of considering the regret are that it is non-negative and that it simplifies asymptotic analysis when T diverges to infinity. However, for any fixed T, any strategy that minimizes the regret also maximizes the sum of expected profits.

the data which suggest that intra-week experimentation might be happening or that the week of monopolists does not coincide with the week defined by NielsenIQ (Thursday-Wednesday). While I would be worried about my time period definition if there was evidence for substantial intra-week pricing, I am not worried about a potential difference in which day is the first in a week. As suggestive evidence, I look at price endings. Empirically, the behavioral left-digit bias, i.e., that consumers tend to perceive a price of $8.99 as much lower than $9, is a powerful force in consumer demand (Strulov-Shlain (2021)). If many firms would engage in intra-week experimentation, we would see many prices endings that are far away from 99 or 98. Figure 1.4 shows that the overwhelming number of price endings is 98 or 99, which suggests that intra-week experimentation is not a first-order concern.[16]

Figure 1.4: The Histogram of Price Endings suggests Limited intra-week Price Experimentation



As discussed in the sample construction, I only consider the 10 whiskeys with highest sales because I cannot observe the posted prices of a liquor that is not sold. Thus firms have to post a $J = 10$-dimensional price vector every week. I make the simplifying assumption that firms do not

[16]While it could be that firms change their prices in multiples of $2 increments only, that appears unlikely.

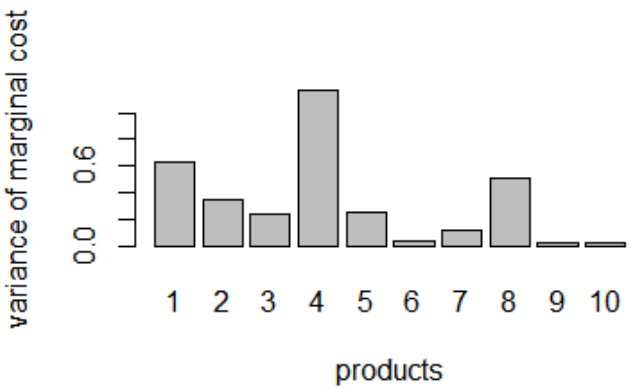
jointly optimize prices. The reason for this is that even if there were only two prices for each product, it would take $2^{10} \approx 1000$ weeks or 20 years to try each combination once. Since we look at $T = 136$ weeks, it seems more reasonable to let firms maximize profits product-by-product rather than jointly. According to the estimates of cross-price elasticities reported in Table 1.5, ignoring price substitution may not be not a dramatic departure from optimality. If the firm knew the shape of demand and the specifics of the logit model, in particular the low-dimensional structure of the substitution patterns, they could learn much more efficiently. But the purpose of this exercise is to imagine that a firm has no information about the structure of demand. So I will simulate the price experimentation product-by-product while keeping the prices of the other 9 products fixed at their average.

A further simplifying assumption in the canonical multi-armed bandit model is stationarity, i.e., that no parameters change over time. In particular, demand, cost, taxes and characteristics are assumed to be constant over the $T = 136$ weeks. This is a strong assumption, in particular for cost and certain characteristics that depend on marketing choices such as whether a product is featured or on display. That marketing variables like whether a product is on feature or display are constant over time is rejected by the data. In principle, it is possible to accommodate such changes by considering contextual multi-armed bandits. But these marketing variables are not really contextual variables. They are choice variables themselves, which are in the control of the store managers. In a future iteration, I would like to allow store managers to also learn about the effects of these marketing variables, simultaneously with prices. I believe that this is feasible even when there are only 136 weeks. Stationarity of demand parameters is a common assumption and taxes are constant. Apart from the marketing variables, I included year and month fixed effects in the demand estimation, but will also average them out to create a stationary demand model. A critical question to assess whether comparing the UCB's performance with the performance of store managers is whether marginal cost for liquor is constant over time. Since we have strong proxies for marginal

cost, we can directly evaluate this in the data. Figure 1.5 illustrates the variance of prices in Oregon over time for each of the 10 whiskeys, suggesting that variation in marginal cost is low for the 10 considered whiskeys in the sample period. In particular, there are some products for which marginal cost are constant.

To apply the UCB algorithm, it is necessary to discretize the space of prices. Theory on multi-armed bandits is silent on how this discretization should be done. However, there are results in economics that can guide us to a judicious choice. Specifically, there is well-established evidence of left-digit bias (Strulov-Shlain (2021)). For a whiskey j, I consider the minimum and maximum price that was ever charged by any firm in my sample. I then take the smallest upper bound that ends on 99 and the largest upper bound that ends on 99 and allow all prices ending on 49 or 99 between those two to define the discrete set of prices. I believe that this is a reasonable discretization, particularly in light of Figure 1.4.

Figure 1.5: Stability of Marginal Cost between June 2012 and Dec 2014



In particular, we see that for whiskey 6, 9 and 10, marginal cost were very stable over time. As a robustness check to departures from stationarity in the cost structure, I will consider these three

whiskeys only.

To conclude, let us formalize the price experimentation problem of monopolist m. For each whiskey $j = 1, \ldots, 10$, the monopolist wants to maximize the sum of expected profits, i.e.,

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left[(p_{j,t}(1-0.17) - mc_j)\text{Bin}(n_m, s_{j,m,t}(p_{j,m,t}))\right] \\ = n_m \sum_{t=1}^{T} (p_{j,t}(1-0.17) - mc_j) s_{j,m,t}(p_{j,m,t}),\end{aligned} \tag{1.29}$$

where the choice probability is given by

$$s_{j,m,t}(p) := \frac{\mathrm{e}^{\bar{x}_j\beta - \alpha P(p,j) + \delta_m + \bar{\xi}_j}}{\mathrm{e}^{\bar{x}_j\beta - \alpha P(p,j) + \delta_m + \bar{\xi}_j} + \sum\limits_{\substack{k=0 \\ k\neq j}}^{J} \mathrm{e}^{\bar{x}_k\beta - \alpha P(\bar{p}_k,k) + \delta_m + \bar{\xi}_k}},$$

$\bar{w}_r$ is a shorthand for $\frac{1}{n}\sum_{t=1}^{t} w_{r,t}$ and $P(p)$ is the effective price paid by the costumer including liquor sales and volume tax, i.e.,

$$P(p, j) := 1.205p + 3.7708\text{liter vol}_j.$$

The choice $p_{j,m,t}$ has to be made with information available up to time t, i.e., all demand realizations between period 1 to $t-1$ but not t.

1.6.6 Experimentation Strategies

I compare four price experimentation strategies by simulating the multi-armed bandit. Specifically, I assume the true demand to follow the simple logit with parameters as reported in Table 1.4. The binomial draws in (1.29) are the source of randomness in the simulation. Since learning strategies are allowed to be functions of the realizations of demand, the path of chosen prices is random, too.

Hence, I simulate an entire experimentation path over $T = 136$ weeks. Three of the four price experimentation algorithms are the UCB algorithm with various concentration parameters. The fourth price experimentation strategy is the one implemented by store managers.

Consider first the price experimentation implemented by store managers. We observe the prices set by store managers, although potentially with some measurement error.[17] This experimentation scheme is independent of the binomial draws in (1.29). Since store managers are not bound by the requirement that their prices be in the discretized set of prices $\{p_{\text{grid},1}, \ldots, p_{\text{grid},L}\}$, I find the closest price in $\{p_{\text{grid},1}, \ldots, p_{\text{grid},L}\}$ to the posted price of every product in every week. The difference cannot be more than a quarter of a dollar due to the construction of the price grid. For most prices, the difference is 0 or 1 penny, as most prices end on .99 or .98, see Figure 1.4.

Now consider the UCB algorithm to experiment with prices for whiskey j.[18] The UCB algorithm will compute finite-sample confidence intervals for profits based on Hoeffding's inequality, recommend to try the price with the highest upper confidence bound for profits and then update the confidence interval for the sampled option. The confidence bounds depend on concentration parameters, $(K_{\text{tail},l})_{l\in\{1,\ldots,L\}}$. I consider the optimal concentration parameters $(K^*_{\text{tail},l})_{l\in\{1,\ldots,L\}}$, estimated concentration parameters $(\hat{K}_{\text{tail},l})_{l\in\{1,\ldots,L\}}$ given the observations made thus far and concentration parameters based on support bounds $(K_{\text{tail, support bound},l})_{l\in\{1,\ldots,L\}}$.

First, consider concentration parameters based on support bounds $(K_{\text{tail, support bound},l})_{l\in\{1,\ldots,L\}}$. These bounds are based on the Remark 8 that I recall here for the reader's convenience.

Remark 8. Consider a real-valued random variable X. If $|X|$ is bounded by B, then X is tail-sub-

[17] See the discussion on intra-week experimentation for a detailed discussion.

[18] I will suppress the dependence on j for notational convenience.

Gaussian with parameter at most $B/\sqrt{\log(2)}$.

The support bound arising from Remark 8 for the profit uses that market shares are bounded by 1. Hence, the profit associated with price p is bounded by

$$B := n(p(1-0.17) - mc), \tag{1.30}$$

where n is the number of consumers, 0.17 is the liquor revenue tax and mc is the marginal cost. Hence,

$$K_{\text{tail, support bound},l} = \frac{n(p_l(1-0.17) - mc)}{\sqrt{\log(2)}}.$$

Second, I consider the estimator for tail-sub-Gaussian parameters derived in section (1.3), i.e.,

$$\hat{K}_{\text{tail},l} = \max_{c=1,\dots,C_l} \frac{\Pi_{c,p_{\text{grid},l}}}{\sqrt{\log\left(\frac{2}{1-\hat{F}(\Pi_{c,p_{\text{grid},l}}-)}\right)}},$$

where C_l counts how often price $p_{\text{grid},l}$ has been explored so far, $\Pi_{c,p_{\text{grid},l}}$ is the observed profit associated with price $p_{\text{grid},l}$ at the c-th try and $\hat{F}$ is the empirical cdf of $(\Pi_{c,p_{\text{grid},l}})_{c=1,\dots,C_l}$.

Third, I consider the optimal concentration parameters. Since I have modeled the demand curve, I can explicitly solve the the smallest possible and hence optimal concentration parameter, i.e.,

$$K^*_{\text{tail},l} := \sup_{t\in\mathbb{R}_+} \frac{t}{\sqrt{\log\left(\frac{2}{1-F_{p_l}(t)}\right)}},$$

where $F_{p_{\text{grid},l}}$ is the true cdf of the profits associated with price $p_{\text{grid},l}$.

1.6.7 Results

I report the results of the comparison of experimentation strategies in Table 1.6. To begin, consider the 'aggregated' estimates for 'all products'. Aggregation refers to adding up the profits of the four local monopoly. We observe that the percentage increase in profits by adopting the UCB algorithm based on support bounds is rather modest with 0.77%. The reason is that concentration parameters based on support bounds are unnecessarily large so that there is too much experimentation and too little exploitation. In contrast, profits would have increased by 26% if the store managers had used the UCB algorithm with estimated concentration parameters based on (1.10). This is a striking increase in profits and close to the 29.9% increase in profits of the infeasible UCB algorithm with optimal concentration parameters.

Table 1.6: Comparison of Price Experimentation Strategies

	UCB(support bound)	UCB($\hat{K}$)	UCB(K^*)
all products			
county 1	3.53%	30.5%	34.7%
county 2	3.56%	29.1%	32.9%
county 3	-6.17%	16.0%	19.2%
county 4	3.60%	29.3%	33.5%
aggregated	0.77%	26.0%	29.9%
products with stable cost			
county 1	- 3.93%	23.51%	25.55%
county 2	- 3.87%	21.35%	23.01%
county 3	-14.52%	7.61%	9.22%
county 4	- 4.99%	20.25%	21.76%
aggregated	- 7.10%	18.26%	20.06%

Notes: The table reports the percentage gain in total profits if store managers had adopted a UCB algorithm between June 2012 to December 2014. The first column is the feasible UCB algorithm with concentration parameters based on support bounds. The second column is the feasible UCB algorithm with estimated concentration parameters based on (1.10)*. The third column is the infeasible UCB algorithm with optimal concentration parameters. The table reports the gains for each local monopoly separately and aggregated. For example, if the monopoly in county 1 had adopted the UCB algorithm with concentration parameters based on support bounds, it would have increased profits for all products by 3.53%. Using the concentration parameters would have increased profits by* 30.5% *and using optimal concentration parameters would have increased profits by* 34.7%*. The 'products with stable cost' are the ones with particularly stable prices over time, see Figure 1.5. The numbers in the table are averages over 10000 simulations.*

These aggregated patterns mask a heterogeneity at the firm level. For example, the price experi-

mentation strategy implemented by monopoly 3 would have outperformed the UCB algorithm with concentration parameters based on support bounds. As a result, the increase in profits achieved by switching to the UCB algorithm with estimated concentration parameters is less pronounced - although a 16% increase in profits is still notable. Huang et al. (2022) also find heterogeneity in the quality/speed of learning at the monopoly level.

One of the main simplifying assumptions to model firm behavior with multi-armed bandits is stability of marginal cost. From Figure 1.4, we have learned that the marginal cost of three products was stable over time. This subset of products allows me to study the sensitivity of results with respect to the assumption of stablility of marginal cost over time. The feasible UCB algorithm with concentration parameters based on support bounds now would have reduced profits compared to the price experimentation implemented by store managers by 7.1%. The feasible UCB algorithm with estimated concentration parameters would still have increased profits by 18.26%, close to 20.06% profit increase achieved by the infeasible UCB with optimal concentration parameters.

Because of the simplifying assumptions to model consumer demand and firm behavior, the comparison between UCB and store managers should be taken with a grain of salt. The three main caveats are non-stationarity of product characteristics, that product interactions have been left aside and that the durability of whiskey was abstracted away. However, the comparison between UCB algorithms is unaffected by these limitations and suggests that (1) the UCB algorithm with estimated concentration parameters performs almost as good as the UCB algorithm with optimal concentration parameters and (2) the UCB algorithm with estimated parameters comfortably outperforms the feasible UCB algorithm with concentration parameters based on support bounds.

1.7 Conclusion

Bandits model learning when agents repeatedly choose actions with uncertain rewards. Bandit models have been applied to experimenting with prices or advertisement to maximize profits, evaluating treatment to maximize outcomes, and testing experience goods to maximize utility. A widely used algorithm for bandit problems is the upper confidence bound (UCB) algorithm. The UCB algorithm builds on Hoeffding's inequality, which involves sub-Gaussian concentration parameters. These parameters are typically not known in applications, constituting an obstacle for applying Hoeffding's inequality for inference.

In this paper, I proposed two estimators for the concentration parameters in Hoeffding's inequality. I studied asymptotic and finite-sample inference based on estimated estimates. With my estimators for concentration parameters, asymptotic inference with estimated parameters is valid under mild conditions and optimal under stronger conditions. Finite-sample inference with estimated parameters is impossible without further assumptions. I proposed conditions under which finite-sample inference with estimated parameters is valid.

These theoretical results can be applied to non-standard inference problems that arise in partial identification and machine learning. One example are linear programs with estimated parameters. I developed a inference method that can account for all parameters to be estimated, and is computationally attractive in interesting settings. In simulations, this approach yields valid confidence sets that are informative but conservative.

Bandit algorithms are the application of my theoretical results on which I focused in this paper. Specifically, I adapted the UCB algorithm to settings in which the sub-Gaussian concentration parameters are not known. Theoretically, I established a finite-sample regret bound for this adapted

UCB algorithm under anti-concentration assumptions. This regret bound is close to the finite-sample regret bound of the UCB algorithm with known sub-Gaussian parameters that leads to the asymptotic optimality of the UCB algorithm. In simulations, I found that UCB with estimated concentration parameter performs almost as well as the UCB algorithm with optimal concentration parameters. In an empirical application on the liberalization of the spirits market in Washington State in 2012, I find that estimating concentration parameters significantly outperforms the available support bounds for concentration parameters.

1.8 Bibliography

Agrawal, R. (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability 27*(4), 1054–1078.

Anderson, T. W. (1969). Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. Technical report, Stanford University.

Angrist, J. D. and W. N. Evans (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 450–477.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association 90*(430), 431–442.

Audibert, J.-Y., R. Munos, and C. Szepesvári (2007). Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, pp. 150–165. Springer.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*(2), 235–256.

Auer, P. and R. Ortner (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica 61*(1-2), 55–65.

Bahadur, R. R. and L. J. Savage (1956). The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics 27*(4), 1115–1122.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2013, 11). Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies 81*(2), 608–650.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pp. 437–478. Springer.

Berry, D. A. (2006). Bayesian clinical trials. *Nature reviews Drug discovery 5*(1), 27–36.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics 37*(4), 1705–1732.

Blundell, R., A. Duncan, and C. Meghir (1998). Estimating labor supply responses using tax reforms. *Econometrica*, 827–861.

Bogoya, J. M., A. Böttcher, and E. A. Maximenko (2016). From convergence in distribution to uniform convergence. *Boletín de la Sociedad Matemática Mexicana 22*(2), 695–710.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Cart. *Classification and Regression Trees*.

Che, Y.-K. and J. Hörner (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics 133*(2), 871–925.

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in airline markets. *Econometrica 77*(6), 1791–1828.

Conlon, C. T. and N. Rao (2015). The price of liquor is too damn high: Alcohol taxation and market structure. *NYU Wagner research paper* (2610118), 2–009.

Crawford, G. S. and M. Shum (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica 73*(4), 1137–1173.

DellaVigna, S. and M. Gentzkow (2019). Uniform pricing in us retail chains. *The Quarterly Journal of Economics 134*(4), 2011–2084.

Dubé, J.-P., A. Hortaçsu, and J. Joo (2021). Random-coefficients logit demand estimation with zero-valued market shares. *Marketing Science 40*(4), 637–660.

Erdem, T. and M. P. Keane (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing science 15*(1), 1–20.

Fang, Z. and A. Santos (2019). Inference on directionally differentiable functions. *The Review of Economic Studies 86*(1), 377–412.

Fang, Z., A. Santos, A. M. Shaikh, and A. Torgovitsky (2022). Inference for large-scale linear systems with known coefficients. *Econometrica, forthcoming*.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences 55*(1), 119–139.

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

Fu, W. and K. Knight (2000). Asymptotics for lasso-type estimators. *The Annals of statistics 28*(5), 1356–1378.

Gilchrist, D. S. and E. G. Sands (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy 124*(5), 1339–1382.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Herstad, E. A. I. (2022). Estimating peer effects with partial network data. *Working paper*.

Hitsch, G. J. and S. Misra (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Volume 1, pp. 278–282. IEEE.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

Honda, J. and A. Takemura (2011). An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning 85*(3), 361–391.

Horowitz, J. L. and S. Lee (2017). Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics 201*(1), 108–126.

Horowitz, J. L. and S. Lee (2022). Inference in a class of optimization problems: Confidence regions and finite sample bounds on errors in coverage probabilities. *Journal of Business & Economic Statistics 0*(0), 1–12.

Hortaçsu, A., J. Lieber, J. Monardo, and A. de Paula (2022). Estimating nesting structures. *Working paper*.

Hortaçsu, A., O. R. Natan, H. Parsley, T. Schwieg, and K. R. Williams (2021). Organizational structure and pricing: Evidence from a large us airline. Technical report, National Bureau of Economic Research.

Hotz, V. J. and R. A. Miller (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies 60*(3), 497–529.

Hsieh, Y.-W., X. Shi, and M. Shum (2022). Inference on estimators defined by mathematical programming. *Journal of Econometrics 226*(2), 248–268.

Huang, Y., P. B. Ellickson, and M. J. Lovett (2022). Learning to set prices. *Journal of Marketing Research 59*(2), 411–434.

Illanes, G. and S. Moshary (2020). Market structure and product assortment: Evidence from a natural experiment in liquor licensure. Technical report, National Bureau of Economic Research.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of political economy 87*(5, Part 1), 972–990.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Keane, M. P. and K. I. Wolpin (1997). The career decisions of young men. *Journal of political Economy 105*(3), 473–522.

Kline, P. and M. Tartari (2016). Bounding the labor supply responses to a randomized welfare experiment: A revealed preference approach. *American Economic Review 106*(4), 972–1014.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 1091–1114.

Lasserre, J. B. (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization 11*(3), 796–817.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms*. Cambridge University Press.

Lieber, J. and T. Wiemann (2022). Demand estimation with finitely many consumers. *Working Paper*.

Mangasarian, O. L. and T.-H. Shiau (1987). Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM Journal on Control and Optimization 25*(3), 583–595.

Maurer, A. and M. Pontil (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics 7*, 1607–1631.

Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political economy 92*(6), 1086–1120.

Misra, K., E. M. Schwartz, and J. Abernethy (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science 38*(2), 226–252.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica 86*(5), 1589–1619.

Mukherjee, S., K. Naveen, N. Sudarsanam, and B. Ravindran (2018). Efficient-ucbv: An almost optimal algorithm using variance estimates. *32*(1).

Nevo, A., J. L. Turner, and J. W. Williams (2016). Usage-based pricing and demand for residential broadband. *Econometrica 84*(2), 411–443.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning 1*(1), 81–106.

Romano, J. P. and M. Wolf (2000). Finite sample nonparametric inference and large sample efficiency. *Annals of Statistics*, 756–778.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review 65*(6), 386.

Rothschild, M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory 9*(2), 185–202.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science 36*(4), 500–522.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of optimization theory and applications 66*(3), 477–487.

Shivaswamy, P. and T. Jebara (2010). Empirical bernstein boosting. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 733–740. JMLR Workshop and Conference Proceedings.

Slawski, M. and M. Hein (2011). Sparse recovery by thresholded non-negative least squares. *Advances in neural information processing systems 24*.

Slawski, M. and M. Hein (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics 7*, 3004–3056.

Strulov-Shlain, A. (2021). More than a penny's worth: Left-digit bias and firm pricing. *Chicago Booth Research Paper* (19-22).

Syrgkanis, V., E. Tamer, and J. Ziani (2021). Inference on auctions with weak assumptions on information. *Working paper*.

Tebaldi, P., A. Torgovitsky, and H. Yang (2019). Nonparametric estimates of demand in the california health insurance exchange. Technical report, National Bureau of Economic Research.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

U.S. Department of Health and Human Services Food and Drug Administration (FDA) (2018). Adaptive designs for clinical trials of drugs and biologics.

Van der Vaart, A. (2000). *Asymptotic statistics*, Volume 3. Cambridge University P ress.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems 4*.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

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

Waisman, C., H. S. Nair, C. Carrion, and N. Xu (2019). Online causal inference for advertising in real-time bidding auctions. *arXiv preprint arXiv:1908.08600*.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, 641–654.

Zhu, Y. (2018). Sparse linear models and l1-regularized 2sls with high-dimensional endogenous regressors and instruments. *Journal of Econometrics 202*(2), 196–213.

1.9 Appendix

1.9.1 On sub-Gaussians

1.9.1.1 Relating Gaussians and sub-Gaussian

While sub-Gaussian inference with estimated concentration parameters may be valid, it is only useful for practitioners if the resulting confidence bounds are "not too large".

Consider X be a normal random variable with mean zero and standard deviation σ. Let us first note that X's MGF is given by

$$\mathbb{E}\left[\exp\left(tX\right)\right] = \exp\left(\frac{t^2\sigma^2}{2}\right).$$

Hence X is MGF-sub-Gaussian with parameter $\frac{\sigma}{\sqrt{2}}$. By Proposition 12, X is tail sub-Gaussian with parameter at most $\sqrt{2}\sigma$.[19] For $\sigma = 1$, this yields a 95% confidence interval $[-2.72, 2.72]$. This has a coverage of 99.34%. This is illustrated in Figure 1.6.

1.9.1.2 Hoeffding's Inequality

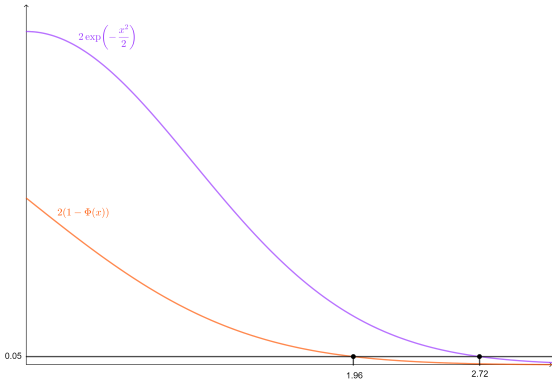
The moment generating function lends itself to study the behavior of sums of independent random MGF-sub-Gaussians.

[19] On the other hand, the tail-sub-Gaussian parameter of the standard normal cannot be smaller than 1.27 since for $t = 4.69494$ we have

$$\mathbb{P}\left[|\mathcal{N}(0,1)| \geqslant t\right] = 2 * (1 - \Phi(t)) \approx 2.66685 * 10^{-6} < 2.321711 * 10^{-6} = 2\exp\left(-\frac{t^2}{1.27^2}\right).$$

Hence the lower bound for the factor from MGF sub-Gaussian parameter to tail sub-Gaussian parameter is $1.27\sqrt{(2)} \approx 1.8$. In Proposition 12, this constant is bounded at 2. So the room for improvement in claim 2 of Proposition 12 is not larger than 10%. Similarly, the factor from tail sub-Gaussian parameter to MGF sub-Gaussian parameter must be bounded from below by 0.557. This is considerably lower than the 1.135441 established in Proposition 12. But this does not imply that the constant in Proposition 12 is inflated since for every fixed distribution, the product of the two costs must be 1. But each step has to hold for *any* sub-Gaussian distribution.

Figure 1.6: Comparison of Gaussian and sub-Gaussian Tail Bounds



Proposition 23. (Vershynin, 2018, Proposition 2.6.1) Let $X_1, \ldots, X_n$ be independent, mean-zero real-valued random variables. Assume that for all i, X_i is an MGF-sub-Gaussian random variable with parameter K_i. Then $\sum_{i=1}^n X_i$ is a MGF-sub-Gaussian random variable with parameter $\sqrt{\sum_{i=1}^n K_i^2}$.

Proof of Theorem 11. Note that by Remark 9, $a_i X_i$ is MGF-sub-Gaussian with parameter $a_i K_i$. By Proposition 23,

$$\sum_{i=1}^n a_i K_i \tag{1.31}$$

is MGF-sub-Gaussian with parameter

$$\sqrt{\sum_{i=1}^n a_i^2 K_i^2}.$$

We can then use Proposition 12 to conclude that (1.31) is tail-sub-Gaussian with parameter

$$2\sqrt{\sum_{i=1}^n a_i^2 K_i^2}.$$

This completes the proof. □

1.9.1.3 Relating Tail- and MGF-sub-Gaussians

Proof of Proposition 12. This proof uses some of the ideas in the proof of Proposition 2.5.2 in Vershynin (2018).

1. Consider a real-valued random variable X that is mean-zero tail sub-Gaussian with parameter $K > 0$. Because of the scaling property 9, we can assume without loss of generality that the K is 1. We have to show that there exists a $\tilde{K}$ such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leqslant \exp\left(\tilde{K}^2\lambda^2\right).$$

In addition to showing existence of such a $\tilde{K}$, the objective is to find the smallest such $\tilde{K}$.[20]

Step 1: MGF of squared random variable

We will first try to bound the MGF of the squared random variable, i.e., $\mathbb{E}\left[\exp\left(\eta^2 X^2\right)\right]$ for any $\eta \in \mathbb{R}$ such that $0 < \eta < 1$. Note that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\eta^2X^2\right)\right] &= \int_0^\infty \mathbb{P}\left[\exp\left(\eta^2X^2\right) > t\right] \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left[\eta^2X^2 > \max\left\{\log\left(t\right), 0\right\}\right] \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left[|X| > \frac{\sqrt{\max\left\{\log\left(t\right), 0\right\}}}{\eta}\right] \mathrm{d}t
\end{aligned}$$

[20] I do not claim to have derived the smallest such constant. Whenever I speak of the "smallest possible" in this proof, I mean the smallest possible that can I can derive given my proof technique.

$$\leqslant \int_0^\infty \min\left\{1, 2\exp\left(-\frac{\max\left\{\log\left(t\right), 0\right\}}{\eta^2}\right)\right\} \mathrm{d}t.$$

The two arguments in the minimum are equal when

$$t = \exp\left(\eta^2 \log\left(2\right)\right).$$

Hence we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\eta^2 X^2\right)\right] \\
&\leqslant \int_0^\infty \min\left\{1, 2\exp\left(-\frac{\log\left(t\right)}{\eta^2}\right)\right\} \mathrm{d}t \\
&= \int_0^{\exp\left(\eta^2\log(2)\right)} 1\mathrm{d}t + 2\int_{\exp(\eta^2\log(2))}^\infty \exp\left(-\frac{\log\left(t\right)}{\eta^2}\right) \mathrm{d}t \\
&= \exp\left(\eta^2\log\left(2\right)\right)\left(1 + \frac{\eta^2}{1-\eta^2}\right) \\
&= \exp\left(\eta^2\log\left(2\right)\right)\left(\frac{1}{1-\eta^2}\right) =: f(\eta).
\end{aligned}$$

Step 2: Small $\boldsymbol{\lambda}$ Consider a threshold parameter $\lambda^* \in (0, 1.3)$[21] that we will choose later. First consider $|\lambda| \leqslant \lambda^*$. Our goal is now to find the smallest α_l such that for all $\lambda \leqslant \lambda^*$

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leqslant \exp(\lambda^2\alpha_l^2).$$

By Lemma 34, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda X\right)\right] &\leqslant \mathbb{E}\left[\lambda X + \exp\left(\kappa\lambda^2 X^2\right)\right] \\
&= \mathbb{E}[\exp\left(\kappa\lambda^2 X^2\right)]
\end{aligned}$$

[21] Note that $1.3\sqrt{\kappa} < 1$.

$$
\begin{aligned}
&\leqslant f(\sqrt{\kappa}\,|\lambda|) && \text{(Step 1)}\\
&\overset{!}{\leqslant} \exp(\lambda^2\alpha_l^2).
\end{aligned}
$$

Hence the smallest α_l is characterized by

$$
\begin{aligned}
\alpha_l^2 &= \sup_{\lambda\leqslant\lambda^*} \frac{\log\left(f(\sqrt{\kappa}\,|\lambda|)\right)}{\lambda^2}\\
&= \sup_{|\lambda|\leqslant\lambda^*} \frac{\log\left(\exp\left(\kappa\lambda^2\log\left(2\right)\right)\left(\frac{1}{1-\kappa\lambda^2}\right)\right)}{\lambda^2}\\
&= \sup_{|\lambda|\leqslant\lambda^*} \frac{\kappa\lambda^2\log\left(2\right)+\log\left(\frac{1}{1-\kappa\lambda^2}\right)}{\lambda^2}\\
&= \log(2)\kappa + \sup_{|\lambda|\leqslant\lambda^*} \frac{\log\left(\frac{1}{1-\kappa\lambda^2}\right)}{\lambda^2}\\
&= \log(2)\kappa + \kappa \sup_{|\lambda|\leqslant\lambda^*} \frac{\log\left(\frac{1}{1-\kappa\lambda^2}\right)}{\kappa\lambda^2}\\
&= \log(2)\kappa + \kappa \sup_{0\leqslant\tau\leqslant\kappa(\lambda^*)^2} \frac{\log\left(\frac{1}{1-\tau}\right)}{\tau}\\
&= \log(2)\kappa - \kappa \inf_{0\leqslant\tau\leqslant\kappa(\lambda^*)^2} \frac{\log\left(1-\tau\right)}{\tau}\\
&= \log(2)\kappa - \kappa\frac{\log\left(1-\kappa(\lambda^*)^2\right)}{\kappa(\lambda^*)^2}\\
&= \log(2)\kappa - \frac{\log\left(1-\kappa(\lambda^*)^2\right)}{(\lambda^*)^2}
\end{aligned}
$$

where in the penultimate step, I used that the function to be minimized is monotonically decreasing.

Step 3: Large $\boldsymbol{\lambda}$

Now consider $|\lambda| \geqslant \lambda^*$. Our goal is to find the smallest α_h such that for all $\lambda \geqslant \lambda^*$

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leqslant \exp\left(\lambda^2 \alpha_h^2\right).$$

For a tuning parameter $0 < \gamma < \sqrt{2}$ that will be chosen later, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda X\right)\right] &\leqslant \exp\left(\frac{1}{2\gamma^2}\lambda^2\right) \mathbb{E}\left[\exp\left(\frac{1}{2}\gamma^2 X^2\right)\right] \\
&\leqslant \exp\left(\frac{1}{2\gamma^2}\lambda^2\right) f\left(\frac{\gamma}{\sqrt{2}}\right) && \text{(Step 1)} \\
&\overset{!}{\leqslant} \exp\left(\alpha_h^2 \lambda^2\right).
\end{aligned}$$

Hence α_h is characterized by

$$\begin{aligned}
\alpha_h^2 &= \inf_{\gamma} \sup_{\lambda \geqslant \lambda^*} \frac{\log\left(\exp\left(\frac{1}{2\gamma^2}\lambda^2\right) f\left(\frac{\gamma}{\sqrt{2}}\right)\right)}{\lambda^2} \\
&= \inf_{\gamma} \frac{1}{2\gamma^2} + \sup_{\lambda \geqslant \lambda^*} \frac{\log\left(f\left(\frac{\gamma}{\sqrt{2}}\right)\right)}{\lambda^2} \\
&= \inf_{\gamma} \frac{1}{2\gamma^2} + \sup_{\lambda \geqslant \lambda^*} \frac{\log\left(\exp\left(\frac{\gamma^2}{2}\log\left(2\right)\right)\left(\frac{1}{1-\frac{\gamma^2}{2}}\right)\right)}{\lambda^2} \\
&= \inf_{\gamma} \frac{1}{2\gamma^2} + \frac{\gamma^2}{2}\frac{1}{\left(\lambda^*\right)^2}\log\left(2\right) + \frac{\log\left(\frac{1}{1-\frac{\gamma^2}{2}}\right)}{\left(\lambda^*\right)^2} \\
&= \inf_{\gamma} \frac{1}{2\gamma^2} + \frac{\log(2)}{2}\frac{1}{\left(\lambda^*\right)^2}\gamma^2 - \frac{1}{\left(\lambda^*\right)^2}\log\left(1 - \frac{1}{2}\gamma^2\right) \\
&= \inf_{0<\tilde{\gamma}<2} \frac{1}{2\tilde{\gamma}} + \frac{\log(2)}{2}\frac{1}{\left(\lambda^*\right)^2}\tilde{\gamma} - \frac{1}{\left(\lambda^*\right)^2}\log\left(1 - \frac{1}{2}\tilde{\gamma}\right) =: z(\tilde{\gamma}).
\end{aligned}$$

Note that $z(\cdot)$ is a convex function on $(0, 2)$ as the sum of three convex functions. Hence any

stationary point is a global optimum. The stationary point is characterized[22] as the solution in $(0, 2)$ to

$$-\frac{1}{(\lambda^*)^2}\frac{\log(2)}{2}\gamma^3 + \frac{1}{(\lambda^*)^2}\left(\log(2)+1\right)\gamma^2 + \frac{1}{2}\gamma - 1 = 0. \tag{1.32}$$

Since this is a polynomial of degree 3, an exact formula for the solutions exists.

Step 4: Optimal Threshold Parameter

Now let us choose λ^* in an optimal way by choosing it such that

$$\min_{\lambda^*<1.3} \max\left\{\alpha_l(\lambda^*), \alpha_h(\lambda^*)\right\}.$$

This can be done numerically, yielding

$$\tilde{K} \approx 1.135441.$$

2. This is the proof of Proposition 2.5.2 in Vershynin (2018). Consider a real-valued random variable X that is MGF-sub-Gaussian with parameter $K > 0$. Then for any $\lambda \in \mathbb{R}$,

$$\mathbb{P}\left[X \geqslant t\right] = \mathbb{P}\left[\exp\left(\lambda X\right) \geqslant \exp\left(\lambda t\right)\right]$$

[22]Note that

$$z'(\gamma) := -\frac{1}{2\gamma^2} + \frac{\log(2)}{2}\frac{1}{(\lambda^*)^2} - \frac{1}{(\lambda^*)^2}\frac{1}{1-\frac{1}{2}\gamma}(-0.5) \stackrel{!}{=} 0.$$

Multiplying this with $2\gamma^2(1-0.5\gamma)$ yields

$$-(1-0.5\gamma) + \frac{\log(2)}{(\lambda^*)^2}\gamma^2(1-0.5\gamma) + \frac{1}{(\lambda^*)^2}\gamma^2 = 0.$$

Rearranging yields (1.32).

$$
\begin{aligned}
&\leqslant \exp\left(-\lambda t\right) \mathbb{E}\left[\exp\left(\lambda X\right)\right] && \text{(Markov)} \\
&\leqslant \exp\left(-\lambda t\right) \exp\left(\lambda^2\right) && \text{(Assumption)} \\
&= \exp\left(\lambda^2 - \lambda t\right).
\end{aligned}
$$

This bound can be minimized by setting $\lambda = \frac{t}{2}$ so that

$$
\mathbb{P}\left[X \geqslant t\right] \leqslant \exp\left(-\frac{t^2}{4}\right).
$$

We can now repeat the same argument with $-X$ to obtain $\mathbb{P}\left[-X \geqslant t\right] \leqslant \exp\left(-\frac{t^2}{4}\right)$. Using a union bound, we conclude

$$
\mathbb{P}\left[|X| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{4}\right).
$$

This shows that X is tail-sub-Gaussian with parameter $K = 2$.

This completes the proof. □

1.9.1.4 Examples of Sub-Gaussian Random Variables

Proof for Remark 8.

1. Consider a real-valued random variable X that is bounded by $-\underline{B}$ from below and $\overline{B}$ from above.

   First consider the tail-sub-Gaussian parameter. For this, let $B := \max\{\underline{B}, \overline{B}\}$. Then for all $t \geqslant 0$,

$$
\mathbb{P}\left[|X| \geqslant t\right] \leqslant \begin{cases} 1 & \text{if } t \leqslant B, \\ 0 & \text{if } t > B. \end{cases}
$$

Consider $t \leqslant B$. Then we have to find K such that

$$t \leqslant 2\exp\left(-\frac{B^2}{K^2}\right).$$

The left hand side is maximized for $t = 1$. So let's find K such that

$$1 \leqslant 2\exp\left(-\frac{B^2}{K^2}\right).$$

Rearranging gives $K \leqslant B/\sqrt{\log(2)}$.

The bound on the MGF-sub-Gaussian parameter follows directly from Hoeffding's lemma.

2. Next consider a normally distributed random variable, i.e., $X \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma \geqslant 0$. Then X is MGF-sub-Gaussian with parameter $\sigma/\sqrt{2}$ as can be seen by inspecting the moment generating function of the Gaussian. See section 1.9.1.1 for details.

This completes the proof. □

1.9.2 Details on Estimation of MGF-parameter

Remark 24. For numeric performance, it is best to avoid evaluating the exponential at large values. For this, I recommend to evaluate $\hat{K}_{\text{mgf}}$ as follows

$$\hat{K}_{\text{mgf}} = \max\left\{ \sup_{\lambda\in\mathbb{R}_+} \frac{\sqrt{\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(\lambda\left(\tilde{X}_i - \max_{i=1,\dots,n}\tilde{X}_i\right)\right)\right) + \lambda\max_{i=1,\dots,n}\tilde{X}_i}}{\lambda}, \right.$$
$$\left. \sup_{\lambda\in\mathbb{R}_-} -\frac{\sqrt{\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(\lambda\left(\tilde{X}_i - \min_{i=1,\dots,n}\tilde{X}_i\right)\right)\right) + \lambda\min_{i=1,\dots,n}\tilde{X}_i}}{\lambda} \right\},$$

where $\tilde{X}_i := X_i - \sum_{j=1}^{n} X_j$.

Remark 25. Suppose $\frac{1}{n}\sum_i X_i \neq 0$. Consider the square of the uncentered (1.15) for simplicity. Then by L'Hospital

$$\limsup_{\lambda \to 0} \frac{\log\left(\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i)\right)}{\lambda^2} = \limsup_{\lambda \to 0} \frac{\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i) X_i}{2\lambda \frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i)} = \infty$$

since first fraction tends to infinity since the denominator tends to zero while the numerator tends to $\frac{1}{n}\sum_i X_i \neq 0$.

Remark 26. Suppose $\frac{1}{n}\sum_i X_i = 0$. Consider the square of (1.15) for simplicity. Then by L'Hospital

$$\begin{aligned}
\limsup_{\lambda \to 0} \frac{\log\left(\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i)\right)}{\lambda^2} &= \limsup_{\lambda \to 0} \frac{\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i) X_i}{2\lambda \frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i)} \\
&= \limsup_{\lambda \to 0} \frac{\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i) X_i^2}{2\frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i) + 2\lambda \frac{1}{n}\sum_{i=1}^{n} \exp(\lambda X_i) X_i} = \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} X_i^2.
\end{aligned}$$

1.9.3 Inferred Realizations

Suppose that we would like to estimate the tail-sub-Gaussian parameter of an unobserved random variable ε for which we have an estimate $\hat{\varepsilon}$. An important example are are residuals in a regression. Suppose that the distribution of ε is does not have point masses so that we can use (1.13). Then

$$\max_{i=1,\ldots,n} \frac{\left|\varepsilon_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} \leqslant \max_{i=1,\ldots,n} \frac{\left|\hat{\varepsilon}_{(i)}\right| + \left|\varepsilon_{(i)} - \hat{\varepsilon}_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}}$$

$$
\begin{aligned}
&\leqslant \max_{i=1,\ldots,n} \frac{\left|\hat{\varepsilon}_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} + \max_{i=1,\ldots,n} \frac{\left|\varepsilon_{(i)} - \hat{\varepsilon}_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} \\
&\leqslant \max_{i=1,\ldots,n} \frac{\left|\hat{\varepsilon}_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} + \frac{\left\|\varepsilon - \hat{\varepsilon}\right\|_\infty}{\min\limits_{i=1,\ldots,n} \sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} \\
&= \max_{i=1,\ldots,n} \frac{\left|\hat{\varepsilon}_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} + \frac{\left\|\varepsilon - \hat{\varepsilon}\right\|_\infty}{\sqrt{\log\left(2\right)}}. \qquad (1.33)
\end{aligned}
$$

This is almost a finite-sample Slutsky result: we can relpace ε with $\hat{\varepsilon}$ at the cost of the term $\|\varepsilon - \hat{\varepsilon}\|_\infty / \sqrt{\log(2)}$. Wether this argument is fruitful depends on whether one can derive a meaningful bound on $\|\varepsilon - \hat{\varepsilon}\|_\infty$ for a particular estimator. For example for OLS residuals, it can be shown[23] that with high probability and for some constant C

$$
\left\|\varepsilon - \hat{\varepsilon}\right\|_\infty \leqslant C K_\varepsilon \sqrt{\frac{p}{n}}.
$$

Note that the unknown K_ε appears in this bound. At first sight, this seems to defeat the purpose as K_ε is the term that is to be estimated in the first place. A trick that is sometimes used in integration by parts[24] is useful here: restoration of the original term. We can estimate the unknown term K_ε with the equally unknown term $\max\limits_{i=1,\ldots,n} |\varepsilon_{(i)}| / \sqrt{\log(2/(1-\frac{i-1}{n}))}$ so that we have restored the original term on the left hand side of (1.33). We then rearrange the equation to see that with high probability,

$$
\max_{i=1,\ldots,n} \frac{\left|\varepsilon_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} \leqslant \frac{1}{1 - C\sqrt{\frac{p}{n}}} \max_{i=1,\ldots,n} \frac{\left|\hat{\varepsilon}_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}}.
$$

[23] I leave out some log terms here to streamline the presentation. See section 1.9.4.1 for a self-contained derivation of a bound on OLS residuals based on sub-Gaussianity and a rank assumption.

[24] See Lemma 36 for an example.

In section 1.4, we will see that left hand side estimates K_ε. We now see that with the finite-sample correction term $1/(1 - Cp/n)$, we can use the inferred realizations for the same purpose, at least when $C\sqrt{p/n} < 1$.

1.9.4 Linear Models

1.9.4.1 OLS

Proposition 27. Consider two random variables $\mathbf{X} \in \mathbb{R}^p$ and a real-valued random variable $\varepsilon \in \mathbb{R}$. Fix a parameter $\beta^* \in \mathbb{R}^p$. Then define

$$Y := \mathbf{X}'\beta^* + \varepsilon.$$

Now suppose that we have n i.i.d. draws of $(Y, \mathbf{X}, \varepsilon)$, of which we only observe $(Y, \mathbf{X})$. Denote the observed random variables by $(Y_1, \ldots, Y_n, \mathbf{X}_1, \ldots, \mathbf{X}_n)$. Denote by X the matrix of n rows and p columns which holds $\mathbf{X}_i$ in row i. Assume that

1. ε_1 is independent of $\mathbf{X}_1$,
2. ε_1 is tail-sub-Gaussian with parameter K_ε,
3. $\|\mathbf{X}_1\|_2$ is tail-sub-Gaussian with parameter K_x,
4. for any $\gamma \neq 0$

$$0 < \varphi := \frac{\gamma^t \frac{1}{n} X^t X \gamma}{\|\gamma\|_2^2}.$$

Then for any $\tau > 0$

$$\mathbb{P}\left[\|\varepsilon - \hat{\varepsilon}\|_\infty \geqslant t\right] \leqslant 4np \exp\left(-\frac{t}{4K_\varepsilon^2 \frac{p}{n} \frac{1}{\varphi^2}} \left(\frac{K_x^2 + \frac{1}{\varphi^2}\frac{4p}{n}K_\varepsilon^2}{K_x^2}\right)\right),$$

where $\hat{\varepsilon}$ are the OLS residuals, $\hat{\varepsilon} := (I - X(X'X)^{-1}X')y$.

Proof. For any norm, we have

$$\|\hat{\varepsilon} - \varepsilon\| = \left\| \left(y - X\hat{\beta}\right) - (y - X\beta^*) \right\| = \left\| X\left(\hat{\beta} - \beta^*\right) \right\|. \tag{1.34}$$

In particular, for the supremum norm, (1.34) implies

$$\begin{aligned} \|\hat{\varepsilon} - \varepsilon\|_\infty &= \left\| X\left(\hat{\beta} - \beta^*\right) \right\|_\infty \\ &= \max_{i=1,\dots,n} \left| \mathbf{X}_i' \left(\hat{\beta} - \beta^*\right) \right| \\ &\leqslant \max_{i=1,\dots,n} \|\mathbf{X}_i\|_2 \left\| \hat{\beta} - \beta^* \right\|_2 && \text{(Hölder)} \\ &= \left\| \hat{\beta} - \beta^* \right\|_2 \max_{i=1,\dots,n} \|\mathbf{X}_i\|_2 . \end{aligned}$$

We proceed by separately bounding the terms $\left\| \hat{\beta} - \beta^* \right\|_2$ and $\max_{i=1,\dots,n} \|\mathbf{X}_i\|_2$ in two steps.

Step 1: Bounding $\left\| \hat{\beta} - \beta^* \right\|_2$

We can use the rank assumption to infer

$$\varphi \left\| \hat{\beta} - \beta^* \right\|_2 \leqslant \frac{1}{\sqrt{n}} \left\| X\left(\hat{\beta} - \beta^*\right) \right\|_2 = \frac{1}{\sqrt{n}} \|\hat{\varepsilon} - \varepsilon\|_2 .$$

For the upper bound,

$$\begin{aligned} \|\hat{\varepsilon} - \varepsilon\|_2 &= \left\| X\left(\hat{\beta} - \beta^*\right) \right\|_2 \\ &= \left\| X\left((X'X)^{-1} X' (X\beta^* + \varepsilon) - \beta^* \right) \right\|_2 \\ &= \left\| \underbrace{X (X'X)^{-1} X'}_{=:P_X} \varepsilon \right\|_2 . \end{aligned}$$

P_X is a projection onto the column space generated by X so that it has the eigenvalue 1 with multiplicity p and the eigenvalue 0 with multiplicity $n - p$ because we assume that X has full rank. Because P_X is also symmetric, it admits an orthogonal diagonalization, i.e., there exists an orthogonal matrix V such that $P_S = VDV^t$ where D is a diagonal matrix whose first p entries on the diagonal are 1 and whose last $n - p$ entries are 0. Column j of matrix V is the eigenvector associated with the j-th element on the diagonal of D. In particular, the first p columns of V are the normalized columns of X because P_X is the projection on the column space of X, so that the eigenvectors corresponding to the eigenvalues 1 are just the normalized columns of X because the column vectors of X trivially span the column space of X. Then

$$
\begin{aligned}
&= \left\| P_X \varepsilon \right\|_2 \\
&= \left\| V^t D V \varepsilon \right\|_2 && \text{(Spectral decomp (P_X is real and symmetric))} \\
&= \left\| V^t D \varepsilon \right\|_2 && \text{(V orthogonal)} \\
&= \sqrt{\sum_{j=1}^{p} \left(\frac{x_j' \varepsilon}{\|x_j\|} \right)^2} \\
&\leqslant \sqrt{p} \max_{j=1,\ldots,p} \left| \frac{x_j' \varepsilon}{\|x_j\|_2} \right| .
\end{aligned}
$$

For any $j \in \{1, \ldots, p\}$, we can use the independence of X and ε to apply Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.3) conditional on x to obtain

$$
\mathbb{P}\left[\left| \frac{x_j'}{\|x_j\|_2} \varepsilon \right| \geqslant t \;\middle|\; x_j \right] \leqslant 2 \exp\left(- \frac{t^2}{4K_\varepsilon^2 \left\| \frac{x_j}{\|x_j\|_2} \right\|_2^2} \right) = \exp\left(- \frac{t^2}{4K_\varepsilon^2} \right) .
$$

As the right hand side does not depend on x, the inequality also holds unconditionally

$$
\mathbb{P}\left[\left| \frac{x_j'}{\|x_j\|_2} \varepsilon \right| \geqslant t \right] \leqslant 2 \exp\left(- \frac{t^2}{4K_\varepsilon^2} \right) .
$$

Now we can use a union bound to take care of the maximum over j

$$\mathbb{P}\left[\max_{j=1,\dots,p}\left|\frac{x_j'}{\|x_j\|_2}\varepsilon\right| \geqslant t\right] \leqslant \sum_{j=1}^{p}\mathbb{P}\left[\left|\frac{x_j'}{\|x_j\|_2}\varepsilon\right| \geqslant t\right] = 2p\exp\left(-\frac{t^2}{4K_\varepsilon^2}\right).$$

To conclude, we have shown

$$\begin{aligned}\mathbb{P}\left[\left\|\hat{\beta}-\beta^*\right\|_2 \geqslant t\right] &\leqslant \mathbb{P}\left[\frac{1}{\varphi}\frac{1}{\sqrt{n}}\|\hat{\varepsilon}-\varepsilon\|_2 \geqslant t\right] \\ &\leqslant \mathbb{P}\left[\frac{1}{\varphi}\sqrt{\frac{p}{n}}\max_{j=1,\dots,p}\left|\frac{x_j'}{\|x_j\|_2}\varepsilon\right| \geqslant t\right] \\ &\leqslant 2p\exp\left(-\frac{\frac{n}{p}\varphi^2t^2}{4K_\varepsilon^2}\right).\end{aligned}$$

Step 2: Bounding $\max_{i=1,\dots,n}\|\mathbf{X}_i\|_2$

A simple union bound gives for any $t>0$

$$\mathbb{P}\left[\max_{i=1,\dots,n}\|\mathbf{X}_i\|_2 \geqslant t\right] \leqslant \sum_{i=1}^{n}\mathbb{P}\left[\|\mathbf{X}_i\|_2 \geqslant t\right] = 2n\exp\left(-\frac{t^2}{K_x^2}\right).$$

Step 3: Conclude

For any $t>0$, we have

$$\begin{aligned}\mathbb{P}\left[\|\hat{\varepsilon}-\varepsilon\|_\infty \geqslant t\right] &\leqslant \mathbb{P}\left[\left\|\hat{\beta}-\beta^*\right\|_2 \max_{i=1,\dots,n}\|\mathbf{X}_i\|_2 \geqslant t\right] \\ &\leqslant \mathbb{P}\left[\left\|\hat{\beta}-\beta^*\right\|_2 \geqslant \sqrt{t}\right]\mathbb{P}\left[\max_{i=1,\dots,n}\|\mathbf{X}_i\|_2 \geqslant \sqrt{t}\right] \\ &\leqslant 2p\exp\left(-\frac{\frac{n}{p}\varphi^2t}{4K_\varepsilon^2}\right)2n\exp\left(-\frac{t}{K_x^2}\right) \\ &\leqslant 4np\exp\left(-t\left(\frac{\frac{n}{p}\varphi^2}{4K_\varepsilon^2}+\frac{1}{K_x^2}\right)\right)\end{aligned}$$

$$\leqslant 4np \exp\left(-t\left(\frac{\frac{n}{p}\varphi^2 K_x^2 + 4K_\varepsilon^2}{4K_\varepsilon^2 K_x^2}\right)\right).$$

Rearranging yields the claimed result.

□

1.9.4.2 The LASSO

In this section, we revisit the analysis of the prediction error of the LASSO.

Our goal is not to generate new insights. In fact, we follow the discussion of Hastie, Tibshirani and Friedman in section 11 of Hastie et al. (2015) which, to the best of our knowledge, is based on the analysis by Bickel, Ritov and Tsybakov in Bickel et al. (2009). The motivation for including this section is to allow the reader to follow the analysis in one coherent framework. While the literature often considers fixed design matrices and Gaussian errors, we present the results with random matrices and sub-Gaussian errors.

We start with the LASSO objective for a generic first stage $j \in \{1, \ldots, p\}$. We have

$$f(\beta) = \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 . \tag{1.35}$$

Denote the minimizer of this function, i.e., the LASSO estimator by $\hat{\beta}^{\text{LASSO}}$.

Lemma 28. We have

$$\left\|\frac{1}{n}\varepsilon^t X\right\|_\infty \left\|\hat{\beta}^{\text{LASSO}} - \beta^*\right\|_1 + \lambda\left(\|\beta^*\|_1 - \left\|\hat{\beta}^{\text{LASSO}}\right\|_1\right) \geqslant \frac{1}{2n}\left\|X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2. \tag{1.36}$$

Proof. As $\hat{\beta}^{\text{LASSO}}$ is the minimizer of (2.49), we have

$$f\left(\beta^*\right) \geqslant f(\hat{\beta}^{\text{LASSO}}).$$

i.e.,

$$\begin{aligned}
&\frac{1}{2n}\left\|y - X\beta^*\right\|_2^2 + \lambda\left\|\beta^*\right\|_1 \\
\geqslant &\frac{1}{2n}\left\|y - X\hat{\beta}^{\text{LASSO}}\right\|_2^2 + \lambda\left\|\hat{\beta}^{\text{LASSO}}\right\|_1 \\
= &\frac{1}{2n}\left\|y - X\beta^* - X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2 + \lambda\left\|\hat{\beta}^{\text{LASSO}}\right\|_1 \\
= &\frac{1}{2n}\left(\left\|y - X\beta^*\right\|_2^2 - 2\left(y - X\beta^*\right)^t X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right) + \left\|X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2\right) + \lambda\left\|\hat{\beta}^{\text{LASSO}}\right\|_1
\end{aligned}$$

Subtracting $\frac{1}{2n}\left\|y - X\beta^*\right\|_2^2$ from both sides, noting that $y - X\beta^* = X\beta^* + \varepsilon - X\beta^* = \varepsilon$ and rearranging, we find

$$\frac{1}{n}\varepsilon^t X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right) + \lambda\left(\left\|\beta^*\right\|_1 - \left\|\hat{\beta}^{\text{LASSO}}\right\|_1\right) \geqslant \frac{1}{2n}\left\|X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2.$$

Using Hölder's inequality to further bound the left hand side of from above, we find (2.50). □

Lemma 29. If

$$\lambda \geqslant \frac{2}{n}\left\|X^t\varepsilon\right\|_\infty$$

then

$$\frac{1}{n}\left\|X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2 \leqslant 12\lambda\left\|\beta^*\right\|_1.$$

Proof. Let's start with (2.50). Note that the lower bound $\frac{1}{2n}\left\|X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2$ can be further

bounded from below by zero. Then we have with the triangle inequality

$$\begin{aligned}
0 \leqslant & \left\| \frac{1}{n} \varepsilon^t X \right\|_\infty \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \lambda \left(\|\beta^*\|_1 - \left\| \hat{\beta}^{\text{LASSO}} \right\|_1 \right) \\
= & \left(\left\| \frac{1}{n} \varepsilon^t X \right\|_\infty - \lambda \right) \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \lambda \left(\left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \|\beta^*\|_1 - \left\| \hat{\beta}^{\text{LASSO}} \right\|_1 \right) \\
\leqslant & \left(\left\| \frac{1}{n} \varepsilon^t X \right\|_\infty - \lambda \right) \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \lambda \left(\left\| \hat{\beta}^{\text{LASSO}} \right\|_1 + \|\beta^*\|_1 + \|\beta^*\|_1 - \left\| \hat{\beta}^{\text{LASSO}} \right\|_1 \right) \\
= & \left(\left\| \frac{1}{n} \varepsilon^t X \right\|_\infty - \lambda \right) \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + 2\lambda \|\beta^*\|_1 \\
\leqslant & -\frac{1}{2} \lambda \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + 2\lambda \|\beta^*\|_1 = \frac{\lambda}{2} \left(4 \|\beta^*\|_1 - \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 \right).
\end{aligned}$$

Comparing the last estimate with the lower bound 0, we find

$$\left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 \leqslant 4 \|\beta^*\|_1 \tag{1.37}$$

Now let's consider (2.50) again:

$$\begin{aligned}
\frac{1}{2n} \left\| X \left(\hat{\beta}^{\text{LASSO}} - \beta^* \right) \right\|_2^2 \leqslant & \left\| \frac{1}{n} \varepsilon^t X \right\|_\infty \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \lambda \left(\|\beta^*\|_1 - \left\| \hat{\beta}^{\text{LASSO}} \right\|_1 \right) \\
\leqslant & \left\| \frac{1}{n} \varepsilon^t X \right\|_\infty \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \lambda \left(\left\| \beta^* - \hat{\beta}^{\text{LASSO}} \right\|_1 \right) \quad \text{(triangle ineq)} \\
\leqslant & \frac{1}{2} \lambda \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 + \lambda \left(\left\| \beta^* - \hat{\beta}^{\text{LASSO}} \right\|_1 \right) \\
= & \frac{3}{2} \lambda \left\| \hat{\beta}^{\text{LASSO}} - \beta^* \right\|_1 \\
\leqslant & \frac{3}{2} \lambda 4 \|\beta^*\|_1 = 6\lambda \|\beta^*\|_1 .
\end{aligned}$$

Multiplying by 2 gives the desired result. □

So far, all arguments were done without using the distribution of X or ε. This came at a price:

the statement in Lemma 48 is conditional on the inequality

$$\lambda \geqslant \frac{2}{n} \left\| X^t \varepsilon \right\|_\infty$$

which, depending on the realizations of X and ε, may or may not hold. Of course, we are interested in choosing λ large enough to ensure that the event holds with "large" probability. For this, we have to impose additional assumptions.

Assumption 1.

1. The observations are independent and identically distributed over i.
2. X and ε are uncorrelated.

Assumption 2.

1. For each j in $\{1, \ldots, p\}$, X_j is tail-sub-Gaussian with a parameter bounded by ρ_X.
2. ε is tail-sub-Gaussian with parameter ρ_ε.

Lemma 30. Suppose Assumption 3 and 4 are satisfied. Set $t \geqslant 0$ arbitrarily. Then with probability of at least

$$1 - 2 \exp\left(\log(p) - c_B \min\left\{ \frac{t^2}{\rho_X^2 \rho_\varepsilon^2}, \frac{t}{\rho_X \rho_\varepsilon} \right\} n \right)$$

we have

$$\max_{j \in \mathfrak{J}} \frac{1}{n} \left\| X^t \varepsilon \right\|_\infty \leqslant t.$$

Proof. We have

$$\frac{1}{n} \left\| X^t \varepsilon \right\|_\infty = \max_{k \in \{1, \ldots, d\}} \left| \frac{1}{n} \sum_{i=1}^{n} X_{i,k} \varepsilon_i \right|.$$

We note that $X_{i,k}\varepsilon_i$ is independent over i by Assumption 3. Also by Assumption 3, the expectation of $X_{i,k}\varepsilon_i$ is zero. In Assumption 4, we have assumed that $X_{i,k}$ are sub-Gaussian with sub-Gaussian norm at most ρ_X and that ε_i is sub-Gaussian with sub-Gaussian norm at most ρ_ε. We know that the product of two sub-Gaussians is sub-exponential with the sub-exponential norm bounded by the product of the sub-Gaussian norms. Hence we can apply Bernstein's inequality to infer that for any $j \in \{1, \dots, p\}$ we have for any $t \geqslant 0$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,k}\varepsilon_i\right| \geqslant t\right] \leqslant 2\exp\left(-c_B \min\left\{\frac{t^2}{\rho_X^2\rho_\varepsilon^2}, \frac{t}{\rho_X\rho_\varepsilon}\right\} n\right).$$

Using a union bound, we find for any $t \geqslant 0$

$$\begin{aligned}
\mathbb{P}\left[\max_{j\in\{1,\dots,p\}}\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\varepsilon_i\right| \geqslant t\right] =& \mathbb{P}\left[\bigcup_{j\in\{1,\dots,p\}}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\varepsilon_i\right| \geqslant t\right\}\right] \\
\leqslant& \sum_{j\in\{1,\dots,p\}} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\varepsilon_i\right| \geqslant t\right] \\
\leqslant& 2p\exp\left(-c_B \min\left\{\frac{t^2}{\rho_X^2\rho_\varepsilon^2}, \frac{t}{\rho_X\rho_\varepsilon}\right\} n\right) \\
=& 2\exp\left(\log(p) - c_B \min\left\{\frac{t^2}{\rho_X^2\rho_\varepsilon^2}, \frac{t}{\rho_X\rho_\varepsilon}\right\} n\right).
\end{aligned}$$

This completes the proof. □

We now combine the results we have shown so far in the following Corollary.

Corollary 31. Suppose Assumption 3 and 4 are satisfied. Fix some $0 < \tau < \sqrt{\frac{n}{\log(p)}}$ and set

$$\lambda = 2\rho_X\rho_\varepsilon\sqrt{\frac{\log(p)}{n}}\tau$$

Then with probability of at least

$$1 - 2\exp\left(\log(p)\left(1 - c_B\tau^2\right)\right) \tag{1.38}$$

it holds for all $j \in \mathfrak{J}$ that

$$\frac{1}{n}\left\|X\left(\hat{\beta}^{\text{LASSO}} - \beta^*\right)\right\|_2^2 \leqslant 24\rho_X\rho_\varepsilon\sqrt{\frac{\log(p)}{n}}\tau\left\|\beta^*\right\|_1. \tag{1.39}$$

Proof. Combine Lemma 48 and Lemma 49. □

In particular, if n and p tend to infinity such that $\frac{\log(p)}{n}$ tends to zero, we can choose $\tau = \frac{1}{2\sqrt{c_B}}$ and see that with probability converging to one, the LASSO prediction error is bounded by a constant times $\sqrt{\frac{\log(p)}{n}}$. This is the so-called slow rate of the LASSO.

1.9.5 Inference

1.9.5.1 Baseline Inference Result

Proof of Theorem 14. We have

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] \in CI(\bar{X}_n, \hat{K} + \gamma, n, \alpha)\right] \\
=&\mathbb{P}\left[\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] \in CI(\bar{X}_n, \hat{K} + \gamma, n, \alpha)\right\} \cap \left\{\hat{K} + \gamma \geqslant K\right\}\right] \\
&+ \underbrace{\mathbb{P}\left[\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] \in CI(\bar{X}_n, \hat{K} + \gamma, n, \alpha)\right\} \cap \left\{\hat{K} + \gamma < K\right\}\right]}_{\geqslant 0} \\
\geqslant&\mathbb{P}\left[\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] \in CI(\bar{X}_n, \hat{K} + \gamma, n, \alpha)\right\} \cap \left\{\hat{K} + \gamma \geqslant K\right\}\right] \\
\geqslant&\mathbb{P}\left[\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] \in CI(\bar{X}_n, K, n, \alpha)\right\} \cap \left\{\hat{K} + \gamma \geqslant K\right\}\right]
\end{aligned}$$

$$\geqslant \mathbb{P}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \in CI(\bar{X}_n, K, n, \alpha)\right\}\right] + \mathbb{P}\left[\left\{\hat{K} + \gamma \geqslant K\right\}\right] - 1.$$

where in the last step, I used that $\mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \geqslant \mathbb{P}[A] + \mathbb{P}[B] - 1$. $\square$

1.9.5.2 Asymptotic Inference

1.9.5.2.1 Conservative Inference with Tail-sub-Gaussian Parameter

Proof of Theorem 15. Let me drop the tail subscript of $\hat{K}_{\text{tail}}$ and K^*_{tail} as this proof is only concerned with tail-sub-Gaussian parameters. Recall the definition of K^* as

$$K^* = \sup_{t \geqslant 0} \frac{t}{\sqrt{\log\left(\frac{2}{1-F(t)}\right)}}.$$

If $K^* < \varepsilon$, then $\mathbb{P}\left[\hat{K} \leqslant K^* - \varepsilon\right] = 0$ for all $n \in \mathbb{N}$ as $\hat{K}$ is non-negative with probability one. So suppose in the following that $K^* \geqslant \varepsilon$. Let $(t_n)_{n\in\mathbb{N}}$ be a sequence that realizes the supremum in the definition of K^*. Then there exists an $N \in \mathbb{N}$ such that for all $n \geqslant N$ we have

$$K^* - \frac{\varepsilon}{2} < \frac{t_n}{\sqrt{\log\left(\frac{2}{1-F(t_n)}\right)}} \leqslant \frac{t_n}{\sqrt{\log\left(\frac{2}{1-F(t_n-)}\right)}},$$

where the last inequality follows from the fact that $F(t_n-) \leqslant F(t_n)$.[25] Next, there exists[26] a $\gamma > 0$ such that

$$F(t_N) = 1 - \gamma.$$

Then

$$\begin{aligned}
\mathbb{P}\left[\hat{K} < K^* - \varepsilon\right] =& \mathbb{P}\left[\hat{K} < K^* - \frac{\varepsilon}{2} - \frac{\varepsilon}{2}\right] \\
\leqslant& \mathbb{P}\left[\hat{K} < \frac{t_N}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)}} - \frac{\varepsilon}{2}\right] \\
=& \mathbb{P}\left[\sup_{t\geqslant 0} \frac{t}{\sqrt{\log\left(\frac{2}{1-\hat{F}(t-)}\right)}} < \frac{t_N}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)}} - \frac{\varepsilon}{2}\right] \\
\leqslant& \mathbb{P}\left[\frac{t_N}{\sqrt{\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)}} < \frac{t_N}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)}} - \frac{\varepsilon}{2}\right] \qquad \text{(monotonicity of } \mathbb{P}\text{)}
\end{aligned}$$

[25] The inequality $F(t_n-) \leqslant F(t_n)$ can be rearranged to $1 - F(t_n) \leqslant 1 - F(t_n-)$. This implies

$$\frac{2}{1-F(t_n-)} \leqslant \frac{2}{1-F(t_n)}.$$

As applying monotonous functions preserves the inequality, this yields $\sqrt{\log\left(\frac{2}{1-F(t_n-)}\right)} \leqslant \sqrt{\log\left(\frac{2}{1-F(t_n)}\right)}$ which can be rearranged to $\frac{1}{\sqrt{\log\left(\frac{2}{1-F(t_n)}\right)}} \leqslant \frac{1}{\sqrt{\log\left(\frac{2}{1-F(t_n-)}\right)}}$. The claimed inequality then follows from multiplication with t_n.

[26] Suppose that such a $\gamma > 0$ doesn't exist. Then $F(t_N) = 1$. Since t_N is finite, this would mean that

$$0 \leqslant K^* - \varepsilon < K^* - \frac{\varepsilon}{2} < \frac{t_N}{\sqrt{\log\left(\frac{2}{1-F(t_N)}\right)}} = \frac{t_N}{\sqrt{\log\left(\frac{2}{0}\right)}} = \frac{t_N}{\sqrt{\log\left(\infty\right)}} = \frac{t_N}{\infty} = 0$$

which is a contradiction.

$$
\begin{aligned}
=&\mathbb{P}\left[\frac{\varepsilon}{2} < \frac{t_N}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)}} - \frac{t_N}{\sqrt{\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)}}\right] \\
=&\mathbb{P}\left[\frac{\varepsilon}{2} < t_N\left(\frac{\sqrt{\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)} - \sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)}}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)}}\right)\right] \\
\leqslant&\mathbb{P}\left[\frac{\varepsilon}{2} < t_N\left(\frac{L\left|F(t_N-) - \hat{F}(t_N-)\right|}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)}}\right)\right] \qquad (1.40)
\end{aligned}
$$

where in the last step, I applied Lemma 32 and that $\gamma < F(t_N-) < 1-\gamma$ which implies that eventually $\frac{\gamma}{2} < \hat{F}(t_N-) < 1-\frac{\gamma}{2}$. Picking up from (1.40), I note

$$
\begin{aligned}
&\mathbb{P}\left[\hat{K} < K^* - \varepsilon\right] \\
&\leqslant \mathbb{P}\left[\frac{\varepsilon}{2} < t_N\left(\frac{L\left|F(t_N-) - \hat{F}(t_N-)\right|}{\sqrt{\log\left(\frac{2}{1-F(t_N-)}\right)\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)}}\right)\right] \\
&\leqslant \mathbb{P}\left[\frac{\varepsilon}{2} < K^*\left(\frac{L\left|F(t_N-) - \hat{F}(t_N-)\right|}{\sqrt{\log\left(\frac{2}{1-\hat{F}(t_N-)}\right)}}\right)\right] \qquad \text{(Definition of } K^* \text{ with supremum)} \\
&\leqslant \mathbb{P}\left[\sqrt{\log(2)}\frac{\varepsilon}{2LK^*} < \left|F(t_N-) - \hat{F}(t_N-)\right|\right].
\end{aligned}
$$

The convergence of this probability to zero follows immediately from the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality. This completes the proof. □

Lemma 32. Fix a $\varepsilon > 0$. Then the function $f : [\varepsilon, 1-\varepsilon] \to \mathbb{R}$ defined by

$$
f(x) := \sqrt{\log\left(\frac{2}{1-x}\right)}
$$

is Lipschitz continuous.

Proof. Note that the derivative

$$f'(x) = \frac{1}{2\sqrt{\log\left(\frac{2}{1-x}\right)}} \frac{1}{1-x}$$

is continuous on the compact set $[\varepsilon, 1-\varepsilon]$ and hence bounded. A bound on the derivative implies that the original function is Lipschitz continuous. □

1.9.5.2.2 Sharp Inference with Tail-sub-Gaussian Parameter

Proof of Theorem 16. Recall (1.8) and (1.11):

$$K^* = \sup_{t \in \mathbb{R}} \frac{t}{\sqrt{\log\left(\frac{2}{1-F(t)}\right)}} = \sup_{p \in [0,1]} \frac{F^{-1}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}}$$

and

$$\hat{K} = \sup_{t \in \mathbb{R}} \frac{t}{\sqrt{\log\left(\frac{2}{1-\hat{F}(t)}\right)}} = \sup_{p \in [0,1]} \frac{\hat{F}^{-1}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}}.$$

Then

$$\begin{aligned}
\mathbb{P}\left[\hat{K} \geqslant K^* + \varepsilon\right] &= \mathbb{P}\left[\sup_{p \in [0,1]} \frac{\hat{F}^{-1}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}} - \sup_{p \in [0,1]} \frac{F^{-1}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}} \geqslant \varepsilon\right] \\
&\leqslant \mathbb{P}\left[\sup_{p \in [0,1]} \frac{\hat{F}^{-1}(p) - F^{-1}(p)}{\sqrt{\log\left(\frac{2}{1-p}\right)}} \geqslant \varepsilon\right] \\
&\leqslant \mathbb{P}\left[\frac{1}{\sqrt{\log(2)}} \sup_{p \in [0,1]} \left|\hat{F}^{-1}(p) - F^{-1}(p)\right| \geqslant \varepsilon\right].
\end{aligned}$$

If the support is bounded and connected, I can apply Theorem 1.1 of Bogoya et al. (2016) to infer that $\sup_{p\in[0,1]} |\hat{F}^{-1}(p) - F^{-1}(p)|$ converges to zero as n tends to infinity. If the support is finite, then uniform convergence of the empirical quantile function follows from its pointwise convergence. This completes the proof. □

1.9.5.3 Finite-Sample Inference

1.9.5.3.1 Impossibility Theorem

Proof of Theorem 17. Fix $K > 0$, $\delta \in (0, K)$, $\varepsilon \in (0, 1)$ and $n \in \mathbb{N}$. Fix $1 > \gamma > 0$. Then consider the distribution

$$X = \begin{cases} 0 & \text{with probability } 1-\gamma, \\ \sqrt{\log\left(\frac{2}{\gamma}\right)} & \text{with probability } \gamma. \end{cases}$$

Let us first verify that X is 1-tail-sub-Gaussian. For this, it is sufficient to look at the tail of X, given by

$$\mathbb{P}\left[|X| \geqslant t\right] = \begin{cases} 1 & \text{if } t \leqslant 0, \\ \gamma & \text{if } 0 < t \leqslant \sqrt{\log\left(\frac{2}{\gamma}\right)}, \\ 0 & \text{if } t > \sqrt{\log\left(\frac{2}{\gamma}\right)}. \end{cases}$$

The only non-trivial sub-Gaussian tail bound is at $t = \sqrt{\log\left(\frac{2}{\gamma}\right)}$, where we verify that

$$\mathbb{P}\left[|X| \geqslant \sqrt{\log\left(\frac{2}{\gamma}\right)}\right] = \gamma \leqslant 2\exp\left(-\sqrt{\log\left(\frac{2}{\gamma}\right)}^2\right) = \gamma.$$

So X is indeed 1-tail-sub-Gaussian. For any $K > 0$, XK is then K-tail sub-Gaussian. Now let us consider an estimator $\hat{K} = \hat{K}(X_1, \ldots, X_n)$ of K given that satisfies (1.19). Consider the event E that $0 = X_1 = \ldots = X_n$. First, note that

$$\mathbb{P}\left[E\right] = (1-\gamma)^n.$$

In particular, we can select γ as a function of n such that $\mathbb{P}\left[E\right] > 1 - \varepsilon$. Secondly, note that on E, $\hat{K} = 0$. Hence

$$\begin{aligned}
\mathbb{P}\left[\hat{K} \geqslant K - \delta\right] =& \mathbb{P}\left[\hat{K} \geqslant K - \delta \middle| E\right] \mathbb{P}[E] + \mathbb{P}\left[\hat{K} \leqslant K - \delta \middle| E^c\right] \mathbb{P}[E^c] \\
\leqslant& \mathbb{P}\left[0 \geqslant K - \delta | E\right] \mathbb{P}[E] + 1\mathbb{P}[E^c] \\
=& 0\mathbb{P}[E] + \mathbb{P}[E^c] < \varepsilon.
\end{aligned}$$

This completes the proof. □

1.9.5.3.2 Positive Result on Finite-Sample Inference

Proof of Theorem 18.

$$\begin{aligned}
&\mathbb{P}\left[\hat{K} > K^*\kappa\right] \\
&= 1 - \mathbb{P}\left[\hat{K} \leqslant K^*\kappa\right] \\
&= 1 - \mathbb{P}\left[\max_{i \in \{1,\ldots,n\}} \frac{\left|X_{(i)}\right|}{\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}} \leqslant K^*\kappa\right] \\
&= 1 - \mathbb{P}\left[\forall i \in \{1,\ldots,n\} \quad \left|X_{(i)}\right| \leqslant K^* \underbrace{\kappa\sqrt{\log\left(\frac{2}{1-\frac{i-1}{n}}\right)}}_{=w_{i,n}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \mathbb{P}\left[\forall i \in \{1, \dots, n\} \quad \left|X_{(i)}\right| \leqslant w_{i,n}\right] \\
&= 1 - \mathbb{E}\left[\prod_{i=1}^{n} \mathbf{1}_{\left|X_{(i)}\right| \leqslant w_{i,n}}\right] \\
&= 1 - \mathbb{E}\left[\mathbf{1}_{X_{(1)} < \ldots < X_{(n)}} \prod_{i=1}^{n} \mathbf{1}_{\left|X_{(i)}\right| \leqslant w_{i,n}}\right] \\
&= 1 - n! \underbrace{\int_0^{w_{1,n}} \ldots \int_0^{w_{n,n}} \mathbf{1}_{t_1 < \ldots < t_n} f(t_1) \ldots f(t_n) \mathrm{d}t_1 \ldots \mathrm{d}t_n}_{=:\mathfrak{I}_n} \\
&\qquad\qquad \text{((Van der Vaart, 2000, Lemma 13.1 (ii)))} \\
&= 1 - n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!} F^i(w_{n-i+1,n}) \mathfrak{I}_{n-i} \qquad \text{(Lemma 35)}
\end{aligned}
$$

□

1.9.6 Regret Bound

Proof of Theorem 20. Consider the events

$$
E(t) := \left\{\exists l = 1, \dots, L \;\; \hat{K}_l(t) < \kappa K_l]\right\}
$$

for all $t = 1 \dots, T$ where $\hat{K}_l^t$ is the estimate of $\hat{K}_l$ in round t based on (1.10) and Proposition 12. In words, $E(t)$ occurs when one tail sub-Gaussian parameter in period t is underestimated. The proof of Theorem 20 proceeds in three steps. First, we will split up the analysis of regret to the case when no $E(t)$ occur and to the case when at least one of them occurs. Second, we will bound the regret when no $E(t)$ occurs using Theorem 8.1 in (Lattimore and Szepesvári, 2020). Third, we will add a crude regret bound in the case that one of the $E(t)$s occurs. Fourth, we will show that (1.22) and (1.23) imply a strong bound on the probability that one of the $E(t)$s occurs.

Step 1: splitting regret

Note that

$$
\begin{aligned}
R_T &= T\mathbb{E}\left[\Pi^*\right] - \sum_{t=1}^{T}\mathbb{E}\left[\Pi_{\mathrm{l(t)}}\right] \\
&= \mathbb{E}\left[T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right] \\
&= \mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] + \mathbb{E}\left[\left(1 - \mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right)\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right]. \qquad (1.41)
\end{aligned}
$$

Step 2: bounding regret when no $E(t)$ occurs

Note that

$$
1 - \mathbf{1}_{\bigcup_{t=1}^{T} E(t)} = \mathbf{1}_{\bigcap_{t=1}^{T} E(t)^c}.
$$

So $1 - \mathbf{1}_{\bigcup_{t=1}^{T} E(t)}$ is the event that no $E(t)$ occurs. In this event, all sub-Gaussian tail-parameters associated with all distributions over all periods are not underestimated. In particular, the estimated sub-Gaussian tail-parameters are valid for inference. In this case, we can divide the reward distributions by the estimated sub-Gaussian parameter and infer that they are now tail-sub-Gaussian with parameter at most 1. This is the condition required for Theorem 8.1 in (Lattimore and Szepesvári, 2020). So that we can infer

$$
\begin{aligned}
&\mathbb{E}\left[\left(1 - \mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right)\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] \\
=&\mathbb{E}\left[\mathbf{1}_{\bigcap_{t=1}^{T} E(t)^c}\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] \\
\leqslant&\mathbb{E}\left[\mathbf{1}_{\bigcap_{t=1}^{T} E(t)^c}\sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L}\inf_{\varepsilon\in(0,\Delta_l)}\Delta_l\left(1 + \frac{5}{\varepsilon^2} + \frac{2\left(\log(1 + T\log^2(T) + 1)\right)}{(\Delta_l - \varepsilon)^2}\right)\right] \\
=&\mathbb{P}\left[\bigcap_{t=1}^{T} E(t)^c\right]\sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L}\inf_{\varepsilon\in(0,\Delta_l)}\Delta_l\left(1 + \frac{5}{\varepsilon^2} + \frac{2\left(\log(1 + T\log^2(T) + 1)\right)}{(\Delta_l - \varepsilon)^2}\right)
\end{aligned}
$$

$$\leqslant \sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L} \inf_{\varepsilon\in(0,\Delta_l)} \Delta_l \left(1 + \frac{5}{\varepsilon^2} + \frac{2\left(\log(1+T\log^2(T)+1)\right)}{(\Delta_l-\varepsilon)^2}\right). \tag{1.42}$$

Step 3: **bounding regret when at least one $E(t)$ occurs**

Next, consider

$$\mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right].$$

If at least one $E(t)$ occurs, Algorithm 3 underestimates at least one tail-sub-Gaussian parameter. In this case, I resort to a worst-case bound on the regret for this case and bound the probability of this event in the next step to control the expected regret. Note that the regret of any strategy is bounded by the regret of always pulling the worst arm.

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] \\
&\leqslant \mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right]\left\|\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right\|_\infty && \text{(Hölder)} \\
&\leqslant \mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right] T \max_{l=1,\dots,L}\Delta_l && (1.43) \\
&\leqslant \mathbb{P}\left[\bigcup_{t=1}^{T} E(t)\right] T \max_{l=1,\dots,L}\Delta_l. && (1.44)
\end{aligned}$$

Step 4: bounding the probability that some $E(t)$ occurs

Using a union bound, we see

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{t=1}^{T} E(t)\right] =& \mathbb{P}\left[\exists l, t : \hat{K}_l < \kappa K_l \text{ in period } t\right] \\
\leqslant& (T-L-1)\frac{1}{T^2}\frac{1}{T-L-1} = \frac{1}{T^2},
\end{aligned} \tag{1.45}$$

where I used that Algorithm 3 estimates exactly $T - L1$ sub-Gaussian tail parameters: after trying each arm twice (which takes $2L$ periods), the Algorithm 3 estimates L sub-Gaussian parameters. Then in each of the following $T-2L-1$ periods, Algorithm 3 estimates 1 tail-sub-Gaussian parameter. In the last period T, there is no need to update the estimate of the tail-sub-Gaussian parameter as there is no need to inform the choice in period $T+1$. This leads to $L+T-2L-1 = T-L-1$ sub-Gaussian parameters. To bound the probability that an underestimating a tail-sub-Gaussian parameter, I used (1.22) and (1.23).

Finally, let me combine the results of the four steps:

$$
\begin{aligned}
R_T =& \mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] + \mathbb{E}\left[\left(1-\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right)\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] \\
& && \text{(using (1.41))} \\
\leqslant & \sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L} \inf_{\varepsilon\in(0,\Delta_l)} \Delta_l\left(1+\frac{5}{\varepsilon^2}+\frac{2\left(\log(1+T\log^2(T)+1)\right)}{(\Delta_l-\varepsilon)^2}\right) \\
& + \mathbb{E}\left[\left(1-\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right)\left(T\Pi^* - \sum_{t=1}^{T}\Pi_{\mathrm{l(t)}}\right)\right] && \text{(using (1.42))} \\
\leqslant & \sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L} \inf_{\varepsilon\in(0,\Delta_l)} \Delta_l\left(1+\frac{5}{\varepsilon^2}+\frac{2\left(\log(1+T\log^2(T)+1)\right)}{(\Delta_l-\varepsilon)^2}\right) \\
& + \mathbb{E}\left[\mathbf{1}_{\bigcup_{t=1}^{T} E(t)}\right] T \max_{l=1,\dots,L}\Delta_l && \text{(using (1.44))} \\
\leqslant & \sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L} \inf_{\varepsilon\in(0,\Delta_l)} \Delta_l\left(1+\frac{5}{\varepsilon^2}+\frac{2\left(\log(1+T\log^2(T)+1)\right)}{(\Delta_l-\varepsilon)^2}\right) \\
& + \frac{1}{T^2} T \max_{l=1,\dots,L}\Delta_l && \text{(using (1.45))} \\
= & \sum_{\substack{l=1 \\ l:\Delta_l>0}}^{L} \inf_{\varepsilon\in(0,\Delta_l)} \Delta_l\left(1+\frac{5}{\varepsilon^2}+\frac{2\left(\log(1+T\log^2(T)+1)\right)}{(\Delta_l-\varepsilon)^2}\right) + \frac{1}{T}\max_{l=1,\dots,L}\Delta_l,
\end{aligned}
$$

as claimed. □

1.9.7 Inference for Linear Programs

Proof of Theorem 22. Consider the class of quadratically constrained quadratic programs, i.e.,

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}x^{\mathrm{T}}P_0x + q_0^{\mathrm{T}}x && (1.46)\\ &\text{subject to} && \tfrac{1}{2}x^{\mathrm{T}}P_ix + q_i^{\mathrm{T}}x + r_i \leqslant 0 \quad \text{for } i = 1,\dots,m, && (1.47)\\ & && Ax = b, && (1.48) \end{aligned}$$

where $P_0, P_1, \dots, P_n \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$ is the optimization variable. It is well known that (1.46) is convex if $P_0, P_1, \dots, P_n$ are positive semidefinite. Now consider the three claims of Theorem 22 in turn.

1. First, when only b is estimated, then $P_0, P_1, \dots, P_n$ are all equal to zero. Hence the quadratically constrained quadratic program reduces to a linear program.

2. Second, when b, c (and potentially A) are estimated, then P_0 is M with zeros to take out $\tilde{b}$ (and potentially $\tilde{A}$). The zeros are inconsequential as they add 0 eigenvalues. The matrix M is the key object to understand the convexity of the program. M is indefinite so that the quadratic program is NP-hard in general. When x and c are constrained to be non-negative, then M is positive semi-definite on the feasible set. In particular, the quadratically constrained quadratic program when A is known is convex in this case as $P_1 = P_2 = \ldots = P_n = 0$. The same argument applies when x and c are known to be non-positive.

3. When A, b, and c are estimated, then the objective matrix P_0 is positive semi-definite provided sign constraints on x and c, as discussed above. Hence the objective is convex. To see

the convexity of the constraints, note that if the pair $(\tilde{A}^1, x^1, \tilde{b}^1)$ and $(\tilde{A}^2, x^2, \tilde{b}^1)$ satisfy

$$\tilde{A}^k x^k \leqslant \tilde{b}^k \tag{1.49}$$

for $k \in \{1, 2\}$, then for any $\lambda \in (0, 1)$,

$$\left(\lambda\tilde{A}^1 + (1-\lambda)\tilde{A}^2\right)\left(\lambda x^1 + (1-\lambda)x^2\right) \leqslant \lambda\tilde{b}_i^1 + (1-\lambda)\tilde{b}_i^2.$$

To see this, expand the left-hand side:

$$\begin{aligned}
&\left(\lambda\tilde{A}^1 + (1-\lambda)\tilde{A}^2\right)\left(\lambda x^1 + (1-\lambda)x^2\right) \\
=&\lambda^2\tilde{A}^1x^1 + \lambda(1-\lambda)\left(\tilde{A}^1x^2 + \tilde{A}^2x^1\right) + (1-\lambda)^2\tilde{A}^2x^2 \\
\leqslant&\lambda^2\tilde{b}^1 + \lambda(1-\lambda)\left(\tilde{A}^1x^2 + \tilde{A}^2x^1\right) + (1-\lambda)^2\tilde{b}^2 && \text{(using (1.49))} \\
\leqslant&\lambda\tilde{b}^1 + \lambda(1-\lambda)\left(\tilde{A}^1x^2 + \tilde{A}^2x^1\right) + (1-\lambda)\tilde{b}^2 && (b^1, b^2 \geqslant 0) \\
\leqslant&\lambda\tilde{b}^1 + (1-\lambda)\tilde{b}^2,
\end{aligned}$$

where in the last step, I used the opposite sign of $\tilde{A}$ and x. All inequalities are understood in a componentwise sense.

This completes the proof. □

1.9.8 Auxiliary Results

1.9.8.1 Uniform Convergence

The first result is to establish uniform convergence in a setting where it is not obvious how the Uniform Law of Large Numbers might be applied. It is a generalization of a well-known result of

real analysis[27] to a stochastic setting. I cannot guarantee that this result is new, but I could not find a reference for it. So, I develop it here.

Proposition 33. Let $a, b \in \mathbb{R}$ such that $a < b$. Consider a function $f : [a, b] \to \mathbb{R}$ which is continuous. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of stochastic processes such that for all n, $f_n : [a, b] \times \Omega \to \mathbb{R}$.[28] Assume further that f_n is

1. monotonically increasing with probability 1 or
2. continuous in its first argument[29] and satisfies

$$\mathbb{P}\left[f_n(\cdot) \text{ is monotonic on } [a, b]\right] \to 1. \tag{1.50}$$

If f_n converges pointwise in probability to f, then f_n converges uniformly in probability to f.

Proof. Fix $\varepsilon > 0$. For all $\delta > 0$, we have to show that there exists an $N \in \mathbb{N}$ such that for all $n \geqslant N$

$$\mathbb{P}\left[\|f_n - f\|_\infty \geqslant \varepsilon\right] < \delta.$$

Since f is continuous on $[a, b]$ and $[a, b]$ is compact, f is uniformly continuous on $[a, b]$. Hence there exists a partition $a = t_1 < \ldots < t_K = b$ of $[a, b]$ such that for all $k \in \{1, \ldots, K - 1\}$,

$$\max_{x\in[t_k,t_{k+1}]} f(x) - \min_{x\in[t_k,t_{k+1}]} f(x) < \frac{1}{2}\varepsilon. \tag{1.51}$$

[27] See, e.g., Proposition 3.2 in Bogoya et al. (2016).

[28] I will suppress the dependence on Ω in the rest of the proof.

[29] The continuity is only needed to ensure the measurability of the event $\{f_n(\cdot) \text{ is monotonic on } [a, b]\}$. For this, note that since f_n is continuous

$$\{f_n(\cdot) \text{ is monotonically increasing on } [a, b]\} = \bigcap_{x,y\in\mathbb{Q}\cap[a,b]:x<y} \{f_n(y) \geqslant f_n(x)\}.$$

Since the event on the left hand side is a countable intersection of measurable sets, it is measurable itself. A similar argument can be made to show that the event $\{f_n(\cdot) \text{ is monotonically decreasing on } [a, b]\}$ is measurable. Then $\{f_n(\cdot) \text{ is monotonic on } [a, b]\}$ can be written as the union of two measurable sets.

Since f_n converges pointwise in probability and $K < \infty$, there exists an $N^1 \in \mathbb{N}$ such that for all $n \geqslant N^1$

$$\mathbb{P}\left[|f_n(t_k) - f(t_k)| \geqslant \frac{1}{2}\varepsilon \quad \forall k = 1, \ldots, K\right] < \frac{1}{2}\frac{\delta}{K-1}. \tag{1.52}$$

By (1.50), there exists an $N^2 \in \mathbb{N}$ such that for all $n \geqslant N^2$,

$$\mathbb{P}\left[f_n(\cdot) \text{ is monotonic on } [a,b]\right] \geqslant 1 - \frac{1}{2}\frac{\delta}{K-1}. \tag{1.53}$$

Then for all $n \geqslant N := \max\{N^1, N^2\}$

$$\begin{aligned}
&\mathbb{P}\left[\|f_n - f\|_\infty \geqslant \varepsilon\right] \\
=&\mathbb{P}\left[\sup_{x\in[a,b]} |f_n(x) - f(x)| \geqslant \varepsilon\right] \\
\leqslant&\sum_{k=1}^{K-1} \mathbb{P}\left[\sup_{x\in[t_k,t_{k+1}]} |f_n(x) - f(x)| \geqslant \varepsilon\right] \\
\leqslant&\frac{1}{2}\delta + \sum_{k=1}^{K-1} \mathbb{P}\left[\max\left\{\sup_{x\in[t_k,t_{k+1}]} |f_n(t_k) - f(x)|, \sup_{x\in[t_k,t_{k+1}]} |f_n(t_{k+1}) - f(x)|\right\} \geqslant \varepsilon\right] \qquad ((1.53)) \\
=&\frac{1}{2}\delta + \sum_{k=1}^{K-1} \mathbb{P}\left[\max\left\{\sup_{x\in[t_k,t_{k+1}]} |f_n(t_k) - f(t_k) + f(t_k) - f(x)|,\right.\right. \\
&\qquad\left.\left.\sup_{x\in[t_k,t_{k+1}]} |f_n(t_{k+1}) - f(t_{k+1}) + f(t_{k+1}) - f(x)|\right\} \geqslant \varepsilon\right] \\
\leqslant&\frac{1}{2}\delta + \sum_{k=1}^{K-1} \mathbb{P}\left[\max\left\{|f_n(t_k) - f(t_k)| + \sup_{x\in[t_k,t_{k+1}]} |f(t_k) - f(x)|,\right.\right. \\
&\qquad\left.\left.|f_n(t_{k+1}) - f(t_{k+1})| + \sup_{x\in[t_k,t_{k+1}]} |f(t_{k+1}) - f(x)|\right\} \geqslant \varepsilon\right] \\
\leqslant&\frac{1}{2}\delta + \sum_{k=1}^{K-1} \mathbb{P}\left[\max\left\{|f_n(t_k) - f(t_k)| + \frac{1}{2}\varepsilon, |f_n(t_{k+1}) - f(t_{k+1})| + \frac{1}{2}\varepsilon\right\} \geqslant \varepsilon\right] \qquad ((1.51)) \\
=&\frac{1}{2}\delta + \sum_{k=1}^{K-1} \mathbb{P}\left[\max\left\{|f_n(t_k) - f(t_k)|, |f_n(t_{k+1}) - f(t_{k+1})|\right\} \geqslant \frac{1}{2}\varepsilon\right]
\end{aligned}$$

$$\leqslant \frac{1}{2}\delta + \sum_{k=1}^{K-1} \frac{1}{2}\frac{\delta}{K-1} \tag{(1.52)}$$

$$= \delta.$$

This completes the proof. □

1.9.8.2 Numerical Inequality

Lemma 34. For all $x \in \mathbb{R}$, we have[30]

$$\exp(x) \leqslant x + \exp\left(0.5575x^2\right).$$

Proof. For notational convenience, define the constant

$$\kappa := 0.5575 \tag{1.54}$$

and the function

$$\begin{aligned} f(x) &= \exp\left(0.5575x^2\right) + x - \exp(x) \\ &= \exp\left(\kappa x^2\right) + x - \exp(x). \end{aligned}$$

We will show that $f(x) \geqslant 0$ for different regions for x.

Case 1: $x \geqslant \frac{1}{\kappa}$

[30] The constant 0.5575 may not be optimal but the inequality is not true if 0.5575 is replaced with 0.5574: try $x = 0.64$.

In this case, we have

$$\kappa x^2 \geqslant x$$

so that $f(x) \geqslant 0$ since $\exp(\cdot)$ is continuous.

Case 2: $0 \leqslant x \leqslant \frac{1}{\kappa}$

Expanding the exponential, we see

$$\begin{aligned}
f(x) &= \sum_{n=0}^{\infty} \frac{(\kappa x^2)^n}{n!} + x - \sum_{n=0}^{\infty} \frac{x^n}{n!} \\
&= 1 + \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} + x - \left(1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}\right) \\
&= \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} - \sum_{n=2}^{\infty} \frac{x^n}{n!} \\
&= \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} - \sum_{n=2}^{\infty} \frac{x^n}{n!} \\
&= (1-\gamma) \sum_{n=0}^{\infty} \gamma^n \kappa x^2 + \sum_{n=2}^{\infty} \frac{1}{n!} \left((\kappa x^2)^n - x^n\right) && \text{(Geom. sum, } \gamma \in (0,1)\text{)} \\
&= (1-\gamma) \sum_{n=2}^{\infty} \gamma^{n-2} \kappa x^2 + \sum_{n=2}^{\infty} \frac{1}{n!} \left((\kappa x^2)^n - x^n\right) \\
&= \sum_{n=2}^{\infty} \frac{1}{n!} \left(n! \gamma^{n-2} (1-\gamma) \kappa x^2 + \kappa^n x^{2n} - x^n\right).
\end{aligned}$$

Now note that it would be sufficient to show that

$$\inf_{x \in \left[0, \frac{1}{\kappa}\right]} f(x) \geqslant 0. \tag{1.55}$$

We have

$$\begin{aligned}
\inf_{x\in\left[0,\frac{1}{\kappa}\right]} f(x) &= \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \left(\sum_{n=2}^{\infty} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 + \kappa^n x^{2n} - x^n \right) \right) \\
&\geqslant \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \left(\sum_{n=2}^{14} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 + \kappa^n x^{2n} - x^n \right) \right) \\
&\quad + \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \left(\sum_{n=15}^{\infty} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 + \kappa^n x^{2n} - x^n \right) \right) \\
&\geqslant \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \left(\sum_{n=2}^{14} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 + \kappa^n x^{2n} - x^n \right) \right) \\
&\quad + \sum_{n=15}^{\infty} \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 + \kappa^n x^{2n} - x^n \right) \\
&\geqslant \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \left(\sum_{n=2}^{14} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 + \kappa^n x^{2n} - x^n \right) \right) \\
&\quad + \sum_{n=15}^{\infty} \inf_{x\in\left[0,\frac{1}{\kappa}\right]} \frac{1}{n!} \left(n!\gamma^{n-2}(1-\gamma)\kappa x^2 - x^n \right).
\end{aligned}$$

Now, notice that the first summand is a polynomial (of order 28). Since finite polynomials can be minimized exactly (Lasserre (2001)), I use software to show that the first term is non-negative for $\gamma = \frac{1}{2}$. So fix $\gamma = 0.5$. For the second term, consider any $n \geqslant 15$. I will show that all summands are non-negative so that the series is well-defined.

$$n!\gamma^{n-2}(1-\gamma)\kappa x^2 - x^n \geqslant 0 \tag{1.56}$$

for $x \in \left(0, \frac{1}{\kappa}\right]$ if

$$x \leqslant \left(n!\gamma^{n-2}(1-\gamma)\kappa \right)^{\frac{1}{n-2}} = \gamma \left(n!(1-\gamma)\kappa \right)^{\frac{1}{n-2}} \leqslant \gamma \left(n! \right)^{\frac{1}{n-2}} \left((1-\gamma)\kappa \right)^{\frac{1}{n-2}}.$$

Now note that $(1-\gamma)\kappa < 1$ so that the sequence $\left((1-\gamma)\kappa\right)^{\frac{1}{n-2}}$ is monotonically increasing in n

and converges to one. In particular,

$$((1-\gamma)\kappa)^{\frac{1}{n-2}} \geqslant ((1-\gamma)\kappa)^{\frac{1}{15-2}} > 0.9.$$

Similarly, $(n!)^{\frac{1}{n-2}}$ is monotonically increasing[31] in n. Hence

$$(n!)^{\frac{1}{n-2}} \geqslant (15!)^{\frac{1}{15-2}} > 8.$$

Hence (1.56) holds when

$$x \leqslant 0.5 \cdot 8 \cdot 0.9 = 3.6.$$

Since $\frac{1}{\kappa} \approx 1.8 < 3.6$, we have shown (1.55).

Case 3: $-\frac{1}{\kappa} \leqslant x \leqslant 0$ We have

$$\begin{aligned}
f(x) =& \sum_{n=0}^{\infty} \frac{(\kappa x^2)^n}{n!} + x - \sum_{n=0}^{\infty} \frac{x^n}{n!} \\
=& 1 + \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} + x - \left(1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}\right) \\
=& \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} - \sum_{n=2}^{\infty} \frac{x^n}{n!} \\
=& \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} - \sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} \frac{x^n}{n!} - \sum_{\substack{n=2 \\ n \text{ odd}}}^{\infty} \frac{x^n}{n!} \\
=& \kappa x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} + \sum_{\substack{n=2 \\ n \text{ odd}}}^{\infty} \frac{|x|^n}{n!} - \sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} \frac{|x|^n}{n!}
\end{aligned}$$

[31] Note that for any n

$$\frac{(n+1)!^{1/(n-1)}}{(n)!^{1/(n-2)}} = (n+1)^{1/(n-1)} n!^{\frac{1}{n-1} - \frac{1}{n-2}} > 1$$

because any root of a number strictly above 1 is strictly above one.

$$
\begin{aligned}
&= \left(\kappa - \frac{1}{2}\right) x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} + \sum_{\substack{n=3 \\ n \text{ odd}}}^{\infty} \frac{|x|^n}{n!} - \sum_{\substack{n=3 \\ n \text{ even}}}^{\infty} \frac{|x|^n}{n!} \\
&= \left(\kappa - \frac{1}{2}\right) x^2 + \sum_{n=2}^{\infty} \frac{(\kappa x^2)^n}{n!} + \sum_{\substack{n=3 \\ n \text{ odd}}}^{\infty} \left(\frac{|x|^n}{n!} - \frac{|x|^{n+1}}{(n+1)!} \right) \\
&= \underbrace{\left(\kappa - \frac{1}{2}\right) x^2}_{\geqslant 0} + \sum_{n=2}^{\infty} \underbrace{\frac{(\kappa x^2)^n}{n!}}_{\geqslant 0} + \sum_{\substack{n=3 \\ n \text{ odd}}}^{\infty} \underbrace{\frac{|x|^n}{n!}}_{\geqslant 0} \underbrace{\left(1 - \frac{|x|}{n+1}\right)}_{\geqslant 0 \text{ for } x \geqslant -\frac{1}{\kappa}}.
\end{aligned}
$$

Case 4: $x < -\frac{1}{\kappa}$

We have

$$
\begin{aligned}
f'(x) &= \exp(\kappa x^2) 2\kappa x + 1 - \exp(x), \\
f''(x) &= \exp(\kappa x^2) 4\kappa^2 x^2 + \exp(\kappa x^2)(2\kappa) - \exp(x) \\
&= \exp(\kappa x^2) \left(4\kappa^2 x^2 + 2\kappa\right) - \exp(x).
\end{aligned}
$$

Hence $f'(-1/\kappa) \approx -11.19$. Since $f''(x) > 0$ for $x < -\frac{1}{\kappa}$, $f'(x) < 0$ for $x < -\frac{1}{\kappa}$. Since $f(-1/\kappa) \approx 4.05 > 0$, we see that $f(x) > 0$ for all $x < -\frac{1}{\kappa}$. □

1.9.8.3 A Recursive Integral

Let $z_{1,n} < \ldots < z_{n,n}$ be a triangular array of real numbers. The purpose of this section is to find a recursive expression for

$$
\mathfrak{I}_{j,n} := \int_0^{z_{1,n}} \ldots \int_0^{z_{j,n}} \mathbf{1}_{t_1 < \ldots < t_j} f(t_1) \ldots f(t_j) dt_j \ldots dt_1
$$

for any $n \in \mathbb{N}, n > 1$ and $j \in \mathbb{N}$ such that $j \leqslant n$, where $\mathfrak{I}_{0,n} := 1$ and $\mathfrak{I}_{1,n} := F(z_{1,n})$, $F : \mathbb{R} \to \mathbb{R}$ is a cdf which admits a density $f : \mathbb{R} \to \mathbb{R}$.

Lemma 35. We have the following recursive expression for $\mathfrak{I}_{j,n}$

$$\mathfrak{I}_{j,n} = \sum_{i=1}^{j} \frac{(-1)^{i+1}}{i!} F^i(z_{j-i+1,n}) \mathfrak{I}_{j-i,n}.^{32} \tag{1.57}$$

Proof. We have

$$\begin{aligned}
&\mathfrak{I}_{j,n} \\
&= \int_0^{z_{1,n}} f(t_1) \int_{t_1}^{z_{2,n}} f(t_2) \dots \underbrace{\int_{t_{j-1}}^{z_{j,n}} f(t_j) dt_j}_{=F(z_{j,n})-F(t_{j-1})} \dots dt_1 && \text{(Tonelli)} \\
&= F(z_{j,n}) \mathfrak{I}_{j-1,n} - \int_0^{z_{1,n}} f(t_1) \int_{t_1}^{z_{2,n}} f(t_2) \dots \underbrace{\int_{t_{j-2}}^{z_{j-1,n}} f(t_{j-1}) F(t_{j-1}) dt_{j-1}}_{\frac{1}{2}(F^2(z_{j-1,n})-F^2(t_{j-2}))} \dots dt_1 && \text{(R36 with } n=1) \\
&= F(z_{j,n}) \mathfrak{I}_{j-1,n} - \frac{1}{2} F(z_{j-1,n}) \mathfrak{I}_{j-2,n} + \frac{1}{2} \int_0^{z_1} f(t_1) \int_{t_1}^{z_2} f(t_2) \dots \underbrace{\int_{t_{j-3}}^{z_{j-2,n}} f(t_{j-2}) F^2(t_{j-2}) dt_{j-2}}_{\frac{1}{3}(F^3(z_{j-2,n})-F^3(t_{j-3}))} \dots dt_1 \\
& && \text{(R36 with } n=2) \\
&= \sum_{i=1}^{j} \frac{(-1)^{i+1}}{i!} F^i(z_{j-i+1,n}) \mathfrak{I}_{j-i,n},
\end{aligned}$$

where in the last step, we used an induction. □

Remark 36. For any cdf F with density f and any integer $n \in \mathbb{N}$, we have

$$\int_a^b f(x) F^n(x) dx = \frac{1}{n+1} \left(F^{n+1}(b) - F^{n+1}(a) \right).^{33}$$

[32] I use the shorthand $F^i(z) := (F(z))^i$.

Proof. Integrate by parts

$$\int_a^b f(x)F^n(x)dx = \left[F(x)F^n(x)\right]_{x=a}^b - \int_a^b F(x)nF^{n-1}(x)f(x)dx$$

and notice that the second summand on the right-hand side is n times the right-hand side. Rearrange and evaluate the integrals. □

1.9.8.4 Basic Inequality

Remark 37. Consider $a, b \in \mathbb{R}$ such that $a > 0$ and $b > 0$. Then

$$\left|\sqrt{a} - \sqrt{b}\right| \leqslant \sqrt{|a-b|}$$

Proof. Note that

$$\left|\sqrt{a} - \sqrt{b}\right|^2 = a - 2\sqrt{ab} + b \leqslant a - 2\min\{a,b\} + b = |a-b|$$

and take the square root. □

[33] I use the shorthand $F^i(x) := (F(x))^i$.

Chapter 2

Estimating Nesting Structures

Coauthored with Ali Hortaçsu*, Julien Monardo†, and Áureo de Paula‡

2.1 Introduction

The nested logit model is commonly used to estimate demand in differentiated products markets, both by researchers and by antitrust practitioners.[1] The nested logit model extends the logit model by grouping products into nests, where products within the same nest will be closer substitutes than products in different nests. The nested logit model is computationally simple – as it can be estimated via a linear regression – and it is consistent with utility maximization by heterogeneous consumers. It has also been generalized to allow for more general nesting structures (i.e., allocations of products into groups). However, one of the main disadvantages of the nested logit model and its generalizations is that the nesting structure must be chosen a priori.

*Kenneth C. Griffin Department of Economics, University of Chicago
†School of Economics, University of Bristol
‡Department of Economics, University College London

[1] See the seminal papers by Goldberg (1995) and Verboven (1996a) as well as Brenkers and Verboven (2006); Björnerstedt and Verboven (2016); Berry et al. (2016); Azar et al. (2019) for more recent papers. See also the Lagardère/Natexis/VUP (2004), TomTom/Tele Atlas (2008), Unilever/Sara Lee (2010) merger cases handled by the European Commission (see CCR - Competition Competence Report Autumn 2013/1) and the Aetna/Humana merger case litigated by the United States District Court for the District of Columbia (see Civil Action No. 16-1494 (JDB)).

In this paper, we propose a novel method to estimate the nesting structure. We rely on a recent generalization of the nested logit model which retains its linear-in-parameter form and allows any possible nesting structure. Specifically, using aggregate data on market shares, prices and product characteristics, we show how using non-negativity constraints coming from economic theory can help estimate the nesting structure by estimating a linear model. Thereby, we are able to obtain substitution patterns than do not depend on a predetermined nesting structure.

The literature has proposed methods to estimate demand models for differentiated products while accounting for the presence of unobserved (by the modeller) product characteristics and dealing with the resulting endogeneity issues (Berry, 1994; Berry et al., 1995; Berry and Haile, 2014). However, demand estimation faces a trade-off between simplicity of estimation and flexibility of the substitution patterns. Indeed, recognizing the limitations of the logit and nested logit models,[2] the literature has proposed demand models that extend the logit model in two different directions. First, since Berry et al. (1995), the standard practice has been to obtain flexible substitution patterns by using the random coefficient logit model, allowing for unobserved consumer heterogeneity in preferences.[3] However, these flexible substitution patterns are obtained at the cost of a complex estimation procedure that requires solving a non-linear, non-convex optimization problem and simulating the demand function.[4]

Alternatively, to accommodate richer substitution patterns, the literature has also proposed

[2]Since the seminal paper by McFadden (1973), the logit model has been the workhorse model for demand estimation purposes. However, the logit model implies that decreases in price of a product reduce the demand for all other products by the same percentage, no matter how similar products are. This restriction is a manifestation of the independence from irrelevant alternatives (IIA) property of the logit model and may lead to counterintuitive conclusions, such as, consumers who buy a BMW being as likely to switch to another luxury car as to a non-luxury car. Furthermore, because of its simple nesting structure, the nested logit model yields restrictive substitution patterns whereby where products within the same nest will be closer substitutes than products in different nests, which may be counter-intuitive in some applications.

[3]The random coefficient logit model was initially developed by Boyd and Mellman (1980) and Cardell and Dunbar (1980). As shown by McFadden and Train (2000), any random utility model can be theoretically approximated by a random coefficient logit model.

[4]This implies handling the associated issues of local optima, choice of starting values, and the accuracy of the simulation (see, e.g., Knittel and Metaxoglou, 2014, and references therein). See Conlon and Gortmaker (2020) for current best practices in the estimation of structural demand models using BLP method. Note that there are other approaches to solve BLP-type problem (Dubé et al., 2012; Lee and Seo, 2015; Salanié and Wolak, 2019).

models that use more general nesting structures. The most prominent examples are specific instances of generalized extreme value (GEV) models developed by McFadden (1978).[5] In particular, the generalized nested logit model generalizes the demand function of the nested logit model by allowing all possible nesting structures (Wen and Koppelman, 2001). However, like the random coefficients logit model, they require solving non-linear, non-convex optimization problems. By contrast, Fosgerau et al. (2021) propose the inverse generalized nested logit (IGNL) model that directly generalizes the inverse demand of the nested logit model by allowing any nesting structure, while retaining its attractive features: the IGNL model is estimated by linear regression and is consistent with the model of heterogeneous, utility-maximizing consumers studied by Allen and Rehbeck (2019).

In any case, both the nesting-based GEV models and the IGNL models require the modeller to define a specific nesting structure before estimation. In some applications, however, there may be no natural nesting structure. For example, in the automobile market, Brenkers and Verboven (2006) use a nested logit model, where products are first allocated to groups according to their market segment (subcompact, compact, standard, intermediate, and luxury) and where groups are then divided into subgroups according to the country of origin (domestic or foreign). By contrast, Grigolon (2020) constructs groups according to an ordering of cars from subcompact to luxury. Determining which of the nesting structures best describes the automobile market is not always obvious.

This paper contributes to the literature by not assuming a specific nesting structure. Instead, we propose to estimate it from aggregate data on market shares, prices and product characteristics using the framework developed by Fosgerau et al. (2021). Specifically, we exploit non-negativity constraints coming from economic theory as well as sparsity constraints to estimate the IGNL model and its nesting structure.

[5] See also the ordered logit (Small, 1987), the product differentiation logit (Bresnahan et al., 1997), the paired combinatorial logit (Koppelman and Wen, 2000), the flexible coefficient multinomial logit (Davis and Schiraldi, 2014), the ordered nested logit (Grigolon, 2020), etc.

We carry out two sets of Monte-Carlo experiments. An application to actual dataset is forthcoming. The first set assesses the empirical performances of our proposed estimator. Simulations show that it performs well in finite sample. In particular, we correctly detect as positive (resp., zero) 96% (resp., 91.78%) of the positive (resp., zero) nesting parameters.

The second set compares our approach to Berry et al. (1995)'s approach (referred to as the BLP method) in terms of implied substitution patterns and markups. Simulations show that our approach performs well in comparison to the BLP approach when it is misspecified. This shows that our approach is able to obtain accurate estimates of the substitution patterns and is thus of great empirical interest given that estimation of demand models for differentiated products is the starting point of many empirical studies.[6]

This paper is linked to two strands of literature. First, it relates to the extensive literature that estimates demand models for differentiated products using Berry (1994)'s and Berry et al. (1995)'s method to handle the endogeneity issues due to the modelling of unobserved product differentiation through the inclusion of unobserved characteristics terms (see e.g. Nevo, 2011; Berry and Haile, 2016; Dubé, 2018, for an overview of the literature). Several papers have proposed nesting-based models in the GEV framework (cf. footnote 7). Horowitz (1987) proposes a statistical test to discriminate among nested logit models. Closest to this paper are Almagro and Manresa (2019) and Aboutaleb et al. (2021) on one hand and Compiani (2020) and Fosgerau et al. (2021) on the other hand. Almagro and Manresa (2019) and Aboutaleb et al. (2021) propose methods to estimate the nesting structure. However, these papers contrast with ours in three respects. First, they rely on the GEV framework, whereas we rely on Fosgerau et al. (2021)'s framework. As a consequence, our estimation procedure involves convex optimization problems, while theirs involve non-convex optimization problems. Second, they do not allow for potentially endogenous, unobserved characteristics terms. Third, they rely on individual-level data, whereas we rely on aggregate data.

[6]Prominent examples of these studies include market power (Berry et al., 1995; Nevo, 2001), new product (Petrin, 2002; Gentzkow, 2007), mergers (Nevo, 2000; Miller and Weinberg, 2017), and taxes and trade policies (Goldberg, 1995; Verboven, 1996a; Berry et al., 1999; Griffith et al., 2019).

Furthermore, observing that, in Berry (1994)'s method, it is the inverse demand function, rather than the demand function, that is the target of estimation, Compiani (2020) and Fosgerau et al. (2021) directly estimate inverse demand functions to obtain substitution patterns and the implied markups. However, Compiani (2020) propose to non-parametrically estimate the inverse demand function, whereas our approach is fully parametric; and Fosgerau et al. (2021) apply their setting by using a model that extends the nested logit model but relies on an assumed nesting structure.

Second, our paper relates to the econometric literature on sparse high-dimensional linear models. For the case of exogenous regressors, the l_1 regularized Dantzig selector and LASSO are natural candidates (Candes et al., 2007; Bickel et al., 2009) Both of these methods involve choosing a regularization parameter. The choice of the tuning parameter in regularized regressions has been recognized as a theoretical challenge in the subsequent literature. More recently, the regularizing potential of non-negativity constraints has been discovered (Slawski and Hein, 2011, 2013; Meinshausen, 2013). By contrast to the methods mentioned above, this does not involve choosing a regularization parameter. For the case of many endogenous regressors, a high-dimensional version of Two-Stage-Least-Squares based on the LASSO has been proposed and studied by Zhu (2018). A Self-Tuning Instrumental Variable estimator based on the Dantzig selector has been proposed and studied by Gautier and Tsybakov (2018) and been further refined using orthogonality with respect to nuisance parameters by Belloni et al. (2017). Our paper adds to this literature by extending the analysis of non-negativity constraints to the case of many endogenous variables.

The remainder of the paper is organized as follows. Section 2.2 introduces the IGNL model and Section 2.3 discusses its estimation and identification. Section 2.4 introduces our estimator. Section 2.5 presents our Monte-Carlo experiments. Section 2.6 concludes.

2.2 The Inverse Generalized Nested Logit Model

Consider a population of consumers who choose from a set of $J+1$ differentiated products, where product $j = 0$ is referred to as the outside good. Each product $j = 1, \ldots, J$ in each market $t = 1, \ldots, T$ is characterized by the vector $(\mathbf{x}_{jt}, \xi_{jt}, p_{jt}, s_{jt})$, where $\mathbf{x}_{jt} \in \mathbb{R}^K$ is a vector of K observed product/market characteristics, $\xi_{jt} \in \mathbb{R}$ is the jt-product/market unobserved characteristics term, $p_{jt} \in \mathbb{R}$ is the price, and $s_{jt} > 0$ is the market share. Following Berry (1994), the product/market unobserved characteristics term represents all product/market characteristics that are unobserved by the researcher but observed by consumers and firms.

Furthermore, assume that each product $j = 1, \ldots, J$ in each market $t = 1, \ldots, T$ is defined by a linear index δ_{jt} (Berry and Haile, 2014) defined by

$$\delta_{jt} = \mathbf{x}_{jt}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt}, \tag{2.1}$$

where α and $\boldsymbol{\beta}$ are parameters to be estimated, and set $\delta_{0t} = 0$ for all $t = 1, \ldots, T$.

Let $\Delta_J^+ \equiv \{(s_0, \ldots, s_J) \in (0, \infty)^{J+1} : \sum_{j=0}^{J} s_j = 1\}$ be the set of non-zero market shares and $\mathbb{R}_0^{J+1} \equiv \{(\delta_0, \ldots, \delta_J \in \mathbb{R}^{J+1} : \delta_0 = 0\}$ be the set of normalized indexes. The behavior of the consumers is described by the inverse demand function

$$\boldsymbol{\sigma}^{-1} = \left(\sigma_0^{-1}, \ldots, \sigma_J^{-1}\right)^{-1} : \Delta_J^+ \to \mathbb{R}_0^{J+1}, \tag{2.2}$$

which, for each market $t = 1, \ldots, T$, gives the vector of product indexes $\boldsymbol{\delta}_t \equiv (\delta_{0t}, \ldots, \delta_{Jt})$ as a function of the vector of nonzero market shares $\mathbf{s}_t \equiv (s_{0t}, \ldots, s_{Jt})$ and some parameter vector $\boldsymbol{\mu}$ to be estimated,

$$\sigma_j^{-1}(\mathbf{s}_t; \boldsymbol{\mu}) = \delta_{jt}, \quad j = 1, \ldots, J, \quad t = 1, \ldots, T. \tag{2.3}$$

The logit and nested logit models are specific instances of the inverse demand model defined in

Equation (2.2) (Berry, 1994). Indeed, the logit model is defined by the following inverse demand equations

$$\sigma_j^{-1}(\mathbf{s}_t) = \ln(s_{jt}) - \ln(s_{0t}) = \delta_{jt}. \tag{2.4}$$

Furthermore, assuming that the choice set is partitioned into nests and the outside good is the only member of its nest, then for a product j in nest g, the nested logit model generalizes the logit model as follows

$$\sigma_j^{-1}(\mathbf{s}_t; \mu) = (1-\mu)\ln(s_{jt}) + \mu \ln\left(\sum_{k \in g} s_{kt}\right) - \ln(s_{0t}) = \delta_{jt}. \tag{2.5}$$

In this paper, we consider the inverse generalized nested logit (IGNL) model developed by Fosgerau et al. (2021), which generalizes the inverse demand function of the nested logit model to allow for any possible nesting structure. Let $\mathcal{G}(j)$ be the set of all groups containing product j and, as for the nested logit model, assume that the outside good is the only member of its group. Then, the IGNL model is defined by

$$\sigma_j^{-1}(\mathbf{s}_t; \boldsymbol{\mu}) = \left(1 - \sum_{g \in \mathcal{G}(j)} \mu_g\right)\ln(s_{jt}) + \sum_{g \in \mathcal{G}(j)} \mu_g \ln\left(\sum_{k \in g} s_{kt}\right) - \ln(s_{0t}) = \delta_{jt}, \tag{2.6}$$

where the vector of nesting parameters $\boldsymbol{\mu} \equiv ((\mu_j)_{j \in \mathcal{J}}, (\mu_g)_{g \in \mathcal{G}})$ satisfies the following assumptions.

[(A1)] $\sum_{g \in \mathcal{G}(j)} \mu_g < 1$ for all $j = 1, \ldots, J$, $\mu_g \geqslant 0$ for all $g \in \mathcal{G}$.

Several comments are in order. First, under Assumptions (A1) – (A2), the IGNL model is invertible, meaning that any observed vector of market shares $\mathbf{s}_t$ can be rationalized by a unique vector of product indexes $\boldsymbol{\delta}_t \in \mathbb{R}_0^{J+1}$.

Second, Assumptions (A1) – (A2), the IGNL model is consistent with a specific instance of the large class of heterogeneous, utility-maximizing consumers studied by Allen and Rehbeck (2019), where the μ_g's control for the distribution of preferences in the population of consumers. This

implies that the parameters μ_g govern substitution between products.

Third, observe that the IGNL model reduces to the logit model when all μ_g equal zero. This means, as for the nested logit model, that the IGNL model allows deviation from independence from irrelevant alternatives (IIA) thanks to its nesting parameters.

Fourth, rearranging Equations (2.6) shows that the IGNL model boils down to the following linear model[7]

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = \mathbf{x}_{jt}\boldsymbol{\beta} - \alpha p_{jt} + \sum_{g\in\mathcal{G}(j)} \mu_g \ln\left(\frac{s_{jt}}{\sum_{k\in g} s_{kt}}\right) + \xi_{jt}. \tag{2.7}$$

However, as highlighted in the following section, the IGNL model has many parameters to be estimated, which implies that standard linear econometric methods may not work. Besides, there are also many endogeneous variables, which requires us to find many valid instruments. In the remainder of this paper, we provide a method to overcome this issue and then conveniently recover the nesting structure from data.

2.3 Estimation

To better understand our main estimating equation (2.7), we write it in matrix form:

$$\underbrace{\begin{pmatrix} \ln\left(\frac{s_{11}}{s_{01}}\right) \\ \ln\left(\frac{s_{21}}{s_{01}}\right) \\ \vdots \\ \ln\left(\frac{s_{JT}}{s_{0T}}\right) \end{pmatrix}}_{=:\mathbf{y}} = \underbrace{\begin{pmatrix} x_{11}^1 & \dots & x_{11}^K \\ x_{21}^1 & \dots & x_{21}^K \\ \vdots & & \vdots \\ x_{JT}^1 & \dots & x_{,T}^K \end{pmatrix}}_{=:\mathbf{X}} \boldsymbol{\beta} - \alpha \underbrace{\begin{pmatrix} p_{,1} \\ p_{21} \\ \vdots \\ p_{JT} \end{pmatrix}}_{=:\mathbf{p}} + \underbrace{\begin{pmatrix} \mathfrak{g}_{11}^{g_1} & \dots & \mathfrak{g}_{11}^{g_G} \\ \mathfrak{g}_{21}^{g_1} & \dots & \mathfrak{g}_{21}^{g_G} \\ \vdots & & \vdots \\ \mathfrak{g}_{JT}^{g_1} & \dots & \mathfrak{g}_{JT}^{g_G} \end{pmatrix}}_{=:\mathfrak{G}} \boldsymbol{\mu} + \underbrace{\begin{pmatrix} \xi_{11} \\ \xi_{21} \\ \vdots \\ \xi_{JT} \end{pmatrix}}_{=:\boldsymbol{\xi}},$$

where we recall that J is the number of products, T is the number of markets, K is the number of product characteristics, $G := 2^J - J - 1$ is the number of nests which are denoted by $g_1, \dots, g_G$

[7] Based on Berry (1994), this was expected because the IGNL model is an inverse demand model that is in closed-form and linear-in-parameters.

and

$$\mathfrak{g}_{jt}^{g} := \ln\left(\frac{s_{jt}}{\sum\limits_{k\in g} s_{kt}}\right)\mathbf{1}_{j\in g}.$$

Let us get a feeling for the dimensions. In a typical application, there will be few characteristics, say $K \sim 10$. There might be $T \sim 200$ markets when exploiting both space and time variation. If there are $J = 10$ products, there will $n = 2000$ observations and $G \sim 1,000$ nests. If we increase the number of products to $J = 20$, there will be $n = 4000$ observations and $G \sim 1,000,000$ nests. As we see, the number of nests increases exponentially in the number of products. This can easily lead to there being more parameters to estimate than there are observations.

Denote the number of parameters to estimate by $p = K + 1 + G$. If there are more parameters to estimate than observations, i.e. $p > n$, this becomes what is commonly referred to as a "high dimensional" problem. In addition, following the literature, we assume that product characteristics $\mathbf{x}_{jt}$ are exogenous (i.e., to be uncorrelated with ξ_{jt}) and we consider that prices and the group share terms are endogenous. Prices are likely to be endogenous as firms may consider both observed and unobserved characteristics when they set their prices. In the presence of unobserved product characteristics, the group share terms $\mathfrak{g}_{jt}^{g}$ are endogenous by construction. This is because market shares are determined by a full system of equations that depends on the vectors of endogenous prices and of unobserved characteristics, and because consumers may choose products while potentially considering the unobserved characteristics.

How could one ever hope to reliably estimate so many parameters with so few observations? Our key assumption is sparsity: we assume that the number of nests with strictly positive nesting coefficient is "small" compared to the number of observations. If we knew which nests have a positive nesting coefficient and we had valid instruments, we could comfortably use conventional methods. This is what has been done in any research project that has used the nested logit: having specified nests with strictly positive nesting parameter, practitioners could use TSLS (or Maximum

Likelihood) to estimate the parameters which were assumed to be strictly positive. Naturally and notably, researchers have chosen to estimate sparse nesting structures, i.e. only considered nested logit models with much less nests than they had observations. In this sense, sparsity is not a new assumption. What is new is that we do not ask the researcher to know which nests have positive nesting parameters but let the data select these nests.

To estimate such a model, one could penalize the parameters, i.e. use Best Subset Selection or a convex relaxation of it such as the LASSO. However, this involves choosing a penalization parameter. This penalization parameter should be small when there are many non-zero coefficients in the Data Generating Process (DGP). In contrast, when the DGP involves few non-zero parameters, the penalization should be higher. So intuitively, asking a researcher for a penalization parameter to estimate nesting structures means asking for the number of nests with strictly positive nesting parameter. There is no reason to assume that a researcher should know the number of nests with strictly positive nesting parameters.[8]

Instead of relying on penalization, we propose to use a regularization that comes directly from economic theory: the nesting parameters are non-negative. As we will show in theory and in simulations, this is enough to discipline the estimation even in when there are more nesting parameters than observations. Here, we briefly describe how the estimation actually works. We propose to estimate (2.7) in two stages. Suppose, for now, that we have enough instruments. We will discuss them instruments later.

In the first stage, we "predict" the endogenous regressors with the instruments. We do not prescribe a particular method for the first stage. When there are many observations and few parameters, a simple linear regression will work. In a high-dimensional case, when there are more parameters than observations, there must also be more instruments than observations and we have to avoid overfitting in the first stage. So we need some regularization in the first stages, too. Since

[8]There are automated ways of choosing penalization parameters such as cross-validation. However, there are very few theoretical results on how cross-validation affects the quality of the parameter estimates.

there is no economic theory to guide us for regularization, we must rely on statistical methods such as the LASSO. Here, penalization and the choice of a penalization parameter is less problematic because in the first stages, we only care about "predicted values" since these are what matters in the second stage. In contrast, in the second stage, we do care about parameter estimates. For prediction, using penalization in the spirit of Best Subset Selection, LASSO, Elastic Net etc is well understood. Our theoretical analysis will not hinge on a specific first-stage method and rather make an assumption on how well it approximates the true predicted (fitted) values. In the appendix, we show for example that the LASSO naturally satisfies these assumptions with a generic choice of penalization parameters. In simulations, we use Ridge Regression with n-fold (generalized) cross validation for computational convenience.

In the second stage, we regress the outcome on the predicted values from the first stages imposing the non-negativity of the nesting parameters. That is, we solve

$$\min_{\boldsymbol{\beta},\alpha,\boldsymbol{\mu}} \frac{1}{n} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \alpha\hat{\mathbf{p}} - \hat{\mathfrak{G}}\boldsymbol{\mu} \right\|_2^2 \qquad \text{subject to} \qquad \mu_g \geqslant 0 \text{ for all } g = 1, \dots, G, \tag{2.8}$$

where $\hat{\mathbf{p}}$ is the predicted price from the first stage and $\hat{\mathfrak{G}}$ is the predicted group matrix from the first stage. Note that (2.8) is a convex problem so that it can be solved efficiently and with global optimality certificates. Computationally, the challenge is to compute $\hat{\mathfrak{G}}$ because $\mathfrak{G}$ has many columns, i.e. there are many nests. For each column of $\mathfrak{G}$, we have to run one first stage. For example, if $J = 20$ and hence $G \sim 1,000,000$, we will have to run about 1,000,000 first stages with at least 1,000,000 instruments in each first stage. Because each first stage can be run separately, we can parallelize this task. We are presently considering using a third-stage such as the Adaptive LASSO proposed by Zou (2006) or hard thresholding proposed by Slawski and Hein (2011).[9]

[9]The Adaptive LASSO might offer oracle properties, i.e. provide estimates which are as good as if the true model was known in advance. It is applicable here because we can use our NN2SLS estimates are consistent. Hence they can be used to choose a tuning parameter for each regressor in a theory-driven way as shown by Zou (2006) in the low-dimensional asymptotic regime. Thresholding offers to directly obtain finite-sample results on selection accuracy (Slawski and Hein, 2011).

2.3.1 Instruments

Instruments should induce exogenous variation in each of the endogenous variables. As highlighted by Fosgerau et al. (2021), instruments for the IGNL model include conventional instruments. First, price endogeneity can be handled using supply-side instruments such as cost and markup shifters (Berry, 1994; Berry et al., 1995; Berry and Haile, 2014). Cost shifters include the Hausman instruments, i.e., prices in other markets (Hausman et al., 1994; Nevo, 2001). Markup shifters include the BLP-type instruments, i.e., functions of the characteristics of competing products (Berry et al., 1995; Gandhi and Houde, 2020). In particular, Gandhi and Houde (2020) propose the differentiation instruments based on the differences of product characteristics, $d_{ijt}^k = x_{it}^k - x_{jt}^k$ for $i = 1, \ldots, J$, $i \neq j$.

Furthermore, identification of the group share terms requires finding exogenous variations in the relative share of product j within its nests, $\ln(s_{jt}/\sum_{k \in g} s_{kt})$. Intuitively, since the nesting parameters μ_g drive the substitution patterns across products, any variable that can reveal the substitution patterns are therefore good candidates as instruments. Following Verboven (1996b); Gandhi and Houde (2020); Fosgerau et al. (2021), for each group share term $\mathfrak{g}_{jt}^g$, we use differentiation instruments based on the differences of product characteristics with other products of the same group, d_{ijt}^k for $i \neq g$ $i \neq j$.

In the simulations of Section 2.5, we find that cost shifters combined with the differentiation IVs introduced by Gandhi and Houde (2020) work well.

2.3.2 Illustration of Non-Negativity Constraints

To illustrate the power of non-negativity constraints, we abstract from endogeneity issues by setting

$$\boldsymbol{\xi} = \mathbf{0}. \tag{2.9}$$

In doing so, we can focus on the second stage and then compare Non-Negative Least Squares (NNLS) to Ordinary Least Squares (OLS). We consider $J = 4$ (inside) products with $K = 3$ characteristics (including a constant) and hence $G = 2^4 - 4 - 1 = 11$ groups or nests. Hence there are $p = G + K + 1 = 15$ parameters to estimate. We consider two regimes. One is the high-dimensional regime with $T = 3$, where there are more parameters to estimate than observations ($n = T * J = 12$). The other is the low-dimensional regime with $T = 200$, where the number of observations ($n = 800$) far exceeds the number of parameters to estimate. This is an illustration, not a Monte Carlo experiment so that we only simulate one dataset.

Table 2.1 summarizes the experiment. The column labelled "Truth" gives the true parameters, where, e.g., $\mu_{\{1,2\}}$ denotes the nesting parameter associated to the group composed of products 1 and 2. As we see, the simulated model is a nested logit model where all inside products are grouped together: all group coefficients are zero, except the one containing all inside products. So the data generating process exhibits sparsity in the sense that most group coefficients are zero. Comparing the estimates, we see that NNLS outperforms OLS in both low- and high-dimensional regimes. In the high-dimensional regime, the OLS estimates are far from the truth, particularly for $\mu_{\{1,2\}}, \mu_{\{1,3\}}, \mu_{\{2,3\}}, \mu_{\{1,2,3\}}, \mu_{\{2,4\}}, \mu_{\{1,3,4\}}$ and $\mu_{\{1,2,3,4\}}$. This is not surprising as the OLS estimator is not unique in this case.[10] By contrast, the NNLS estimates are relatively close to the truth.

In the low-dimensional regime, we increase the number of observations. As expected, the OLS estimates gain accuracy. However, not all nesting parameters estimates are non-negative, as required by economic theory. In fact, one cannot hope for uniformly non-negative nesting estimates. Since OLS is unbiased, it is likely that some estimates will fall below zero if the true parameter is zero. By contrast, the NNLS estimates are extremely close to the truth.

We also compare NNLS and OLS in terms of estimated price elasticities of demand and markups. In the high-dimensional case, we find that NNLS provides results that are close to the

[10]For the implementation, we use cyclic coordinate descent so that we are guaranteed to converge to a particular minimizer of OLS.

truth, whereas OLS gives biased results. In the low-dimensional case, the two estimators perform equally well.

Table 2.1: Comparing OLS and NNLS for $J = 4$

		$T = 3$		$T = 200$	
	Truth	OLS	NNLS	OLS	NNLS
β_0	0.25	0.137	0.246	0.244	0.250
β_1	0.50	0.525	0.502	0.500	0.500
β_2	0.75	0.738	0.749	0.750	0.750
α	1	1.005	1.000	1.005	1.000
$\mu_{\{1,2\}}$	0	0.530	0.010	−0.021	0
$\mu_{\{1,3\}}$	0	0.489	0.017	−0.021	0
$\mu_{\{2,3\}}$	0	0.591	0.010	−0.021	0
$\mu_{\{1,2,3\}}$	0	−0.370	0	0.045	0.000
$\mu_{\{1,4\}}$	0	0.106	0.001	−0.021	0
$\mu_{\{2,4\}}$	0	0.277	0	−0.021	0
$\mu_{\{1,2,4\}}$	0	0.074	0.017	0.047	0.000
$\mu_{\{3,4\}}$	0	0.013	0	−0.022	0
$\mu_{\{1,3,4\}}$	0	0.236	0.011	0.047	0.000
$\mu_{\{2,3,4\}}$	0	0.074	0.016	0.048	0.000
$\mu_{\{1,2,3,4\}}$	0.5	−0.097	0.462	0.418	0.499
Own-price elasticities	-4.536	-5.876	-4.565		
Cross-price elasticities	0.808	1.254	0.817		
Markups	0.236	0.201	0.235		
Own-price elasticities	-4.485			-4.485	-4.485
Cross-price elasticities	0.778			0.778	0.778
Markups	0.241			0.241	0.241

Notes: The two bottom panels give the estimated own- and cross-price elasticities of demands as well as the estimated (relative) markups, averaged across products and markets.

2.3.3 Illustration of Scaling

As the number of nesting parameters to estimate is exponential in the number of inside products J, there is a curse of dimensionality in estimating nesting structures. No purely data-driven method will be able to fully overcome this curse of dimensionality. It is useful to see how far we can increase the number of products. In particular, it is interesting to see how long it takes to solve the NNLS (second-stage) problem and how good estimates are.

For this illustration, we set the number of markets at $T = 10$ and consider $J = 5, 10, 15, 20$ products. Otherwise, we use the same DGP as in the previous section. In particular, recall that we abstract from endogeneity issues to focus on the second stage. The number of observations is n and the number of parameters to estimate is $p = 2^J - J - 1 + K + 1$. The only nesting parameter in the DGP is the coefficient of the group with all inside products and it has a nesting coefficient of 0.5. Since we implement NNLS via a cyclic coordinate descent, this is a challenging DGP since we first cycle through all other groups. As we see, the number of observations grows linearly in J while the number of parameters to estimate grows exponentially in J. Table 2.2 shows, for all values for J, that NNLS performs well in correctly including the positive nesting parameter (see column "Correctly Included"). It also shows that it falsely includes only a very small percentage of the zero nesting parameters (see column "Falsely Included").

Table 2.2: Scaling Properties of NNLS

			Nesting Parameters			
J	p	n	Correctly Included	Falsely Included	$\|\hat{\mu} - \mu\|_2$	Wall Time
5	30	50	1/1	8	0.006	¡1sec
10	1017	100	1/1	45	0.059	23sec
15	32,756	150	1/1	144	0.492	576sec
20	1,048,555	200	1/1	358	0.513	2991sec

2.4 Econometric Theory

Since this section might be of independent interest, we first present the theory for a general linear model with many endogenous variables. We then show how the theory applies to our problem.

2.4.1 General Theory

We begin with some notation. For a matrix $M \in \mathbb{R}^{k \times l}$ and a set of indices $\mathcal{I} \subset \{1, \ldots, l\}$, the matrix $M_{\mathcal{I}} \in \mathbb{R}^{k \times |\mathcal{I}|}$ corresponds to the submatrix of M which has only the columns of M whose indices are in $\mathcal{I}$. In particular, if $\mathcal{I} = \{i\}$, then $M_{\{i\}}$, usually written M_i, is understood to be the i-th column of M. We write $M_{i,\cdot}$ to designate row i of matrix M,

2.4.1.1 Setting

Suppose we have n independent observations generated as

$$y_i = X_{i,\cdot}\beta^* + \varepsilon_i \tag{2.10}$$

where $y_i \in \mathbb{R}$ is the outcome, $X_{i,\cdot} \in \mathbb{R}^p$ is a rowvector of regressors, $\beta^* \in \mathbb{R}^p$ is a vector of coefficients and $\varepsilon_i \in \mathbb{R}$ is an error term. Stack these row-wise over i so that $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $X = (X_{1,\cdot}, \ldots, X_{n,\cdot}) \in \mathbb{R}^{n \times p}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. X and y are observable but β and ε are unknown. The error term is centered, i.e. $\mathbb{E}[\varepsilon] = 0$. Some (perhaps all) regressors are endogenous, i.e.

$$\mathbb{E}[X_j \varepsilon] \neq 0 \qquad \text{for some } j.$$

There are d instruments $Z_{i,\cdot} \in \mathbb{R}^d$ available so that for each j

$$X_{i,j} = Z_{i,\cdot} \pi_j^* + \eta_{i,j}. \tag{2.11}$$

We assume that instruments Z are independent of first and second stage errors ε and η. The π_j^* have to be estimated for each j, using *some* estimator. The theoretical analysis allows for a class of potential estimators in the first stage as long as they meet our formal assumption 6. It is a strength of the analysis not to specify the first-stage procedure because it allows a practitioner to choose a first stage procedure depending on the DGP at hand. We show in Appendix 2.8.3 that the LASSO satisfies these assumptions with a theoretically guided choice of the tuning parameter.

Denote the first-stage estimates by $\hat{\pi}_j$ for all j. Compute the fitted values as

$$\hat{X}_j := Z \hat{\pi}_j. \tag{2.12}$$

Then the non-negative two stage least squares estimator (NN2SLS) is

$$\hat{\beta}^{\text{NN2SLS}} := \underset{b \in \mathbb{R}^p_+}{\arg\min} \frac{1}{n} \left\| y - \hat{X} \beta \right\|_2^2. \tag{NN2SLS}$$

2.4.1.2 Results

Our main result is a finite-sample concentration result on $\hat{\beta}^{\text{NN2SLS}}$, i.e. that with high probability, $\hat{\beta}^{\text{NN2SLS}}$ is close to β^*.

For this result, we need four assumptions. The first assumption imposes independence of the observations and exogeneity of the instruments. The second assumption is a sub-Gaussian tail assumption on instruments and first and second stage errors. It is similar in spirit to assumptions that first and second moments are bounded but much stronger: all moments exist and cannot grow faster than a certain rate. The third assumption should be thought of as an instrument exogeneity assumption and is explained in detail below. The fourth assumption guarantees a certain quality of first-stage predicted values and is formulated without reference to a particular first-stage method. In section 2.8.3, we show that the LASSO satisfies these assumptions with a theoretically guided choice of the tuning parameter.

Theorem 38. Suppose Assumptions 3, 4, 5 and 6 are satisfied. Then with probability at least

$$p_{T1} := p_{L2} + p_{A3,1,n} + p_{A3,2,n} - 2$$

it holds that

$$\left\| \hat{\beta}^{\text{NN2SLS}} - \beta^* \right\|_1 \leqslant r_{L3} + r_{L2} \max \left\{ \frac{s}{\phi}, \frac{3}{\nu} + \frac{1}{\sqrt{\nu}\sqrt{s}\left(\max_{j \in S} r_{L45}(j) \right)} \right\} =: r_{T1},$$

where p_{L2}, r_{L2} is defined in Lemma 40, r_{L3} is defined in Lemma 41 and r_{L45} is defined in Lemma 45.

To show the bite of theorem 38 without boring the reader with the exact expressions for p_{T1} and r_{T1}, let us also impose the growth conditions formalized in assumption 7. In particular, this entails that n converges to infinity faster than s, $\log(p)$ and $\log(d)$. Then p_{T1} converges to 1 while

r_{T1} converges to zero, i.e. the NN2SLS estimator is consistent in the sense that $\left\|\hat{\beta}^{\text{NN2SLS}} - \beta^*\right\|_1$ converges in probability to zero as n tends to infinity.

Corollary 39. Under Assumptions 3, 4, 5 6 and 7, $\hat{\beta}^{\text{NN2SLS}}$ is l_1-consistent in probability for β^*. The rate of convergence is at least

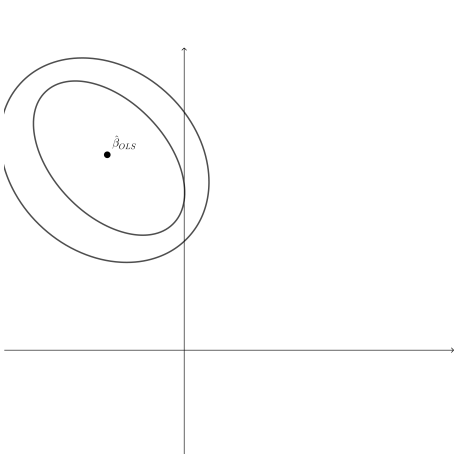
$$\max\left\{\frac{s^3}{n^\gamma}, \frac{s^{\frac{3}{2}}}{n^\alpha}, s^3 r_{A4,S,n}, s^{\frac{3}{2}} r_{A4,\{1,\ldots,p\},n}\right\}. \tag{2.13}$$

We proceed by sketching the proof of theorem 38 and then presenting our assumptions. The actual proofs are in section 2.8.1 of the appendix.

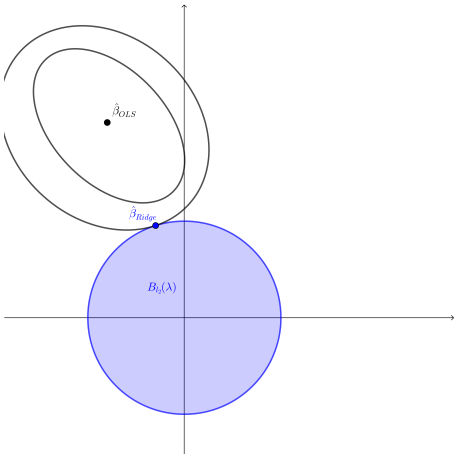
2.4.1.3 Intuition

To get an intuition for how non-negativity constraints can induce sparsity and thereby regularize a high-dimensional linear model, consider figure 2.1. It consists of four panels. Each panel has a coordinate axes on which we consider the parameters β_1, β_2 of a linear regression with two regressors. Panel 2.1a shows the contour lines of the least squares objective and the OLS estimator $\hat{\beta}^{OLS}$ which minimizes the OLS objective. The other panels consider the problem of minimizing the OLS objective under different constraints.

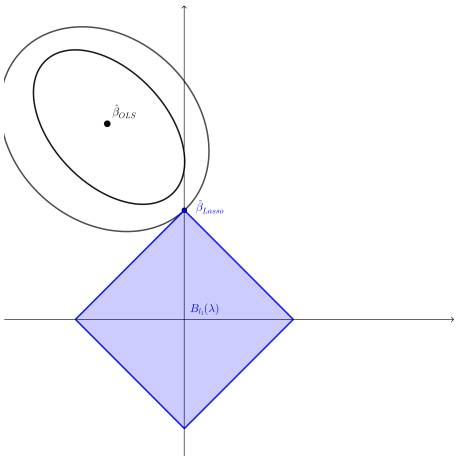
Figure 2.1: Inducing Sparsity without Penalization: Non-negativity Constraints



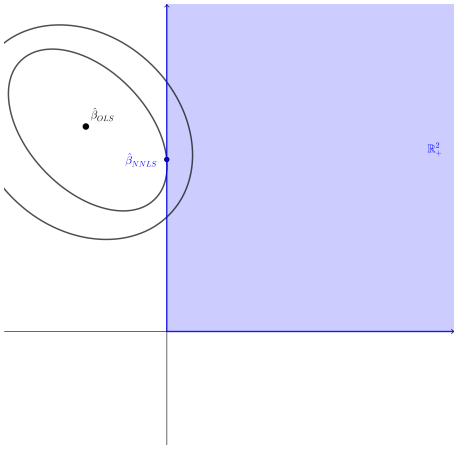
(a) OLS



(b) Ridge



(c) LASSO



(d) NNLS

Panel (b) considers the case of a constraint on the l_2 norm of β, i.e. Ridge regression with a fixed penalization parameter. The penalization parameter is directly related to the radius of the ball that is the feasible set. We see that Ridge minimizes a convex objective over a convex set. The Ridge estimator $\hat{\beta}^{\text{Ridge}}$ is closer to the origin than the OLS estimator but does not induce sparsity: no component of $\hat{\beta}^{\text{Ridge}}$ is exactly zero. Panel (c) visualizes the case of a constraint on the l_1 norm of β, i.e. the LASSO with a fixed penalization parameter, in the spirit of the famous figure 2 of Tibshirani (1996). We see that the feasible set of the LASSO is also feasible. In contrast to the

Ridge estimator, the LASSO can induce sparsity: here $\hat{\beta}_1^{\text{LASSO}} = 0$. Finally, panel (d) visualizes constraints on the sign of beta, i.e. non-negativity constraints. We see that the NNLS estimator $\hat{\beta}^{NNLS}$ can also induce sparsity when components of the OLS estimator are negative. In higher dimensions, the bite of the non-negativity constraints rises as the number of orthants increases exponentially in the dimension of β while we only consider one orthant. The figure also visualizes how NNLS does not depend on choosing a penalization parameter.

2.4.1.4 Sketch of Proof

To understand why the NN2SLS estimator works well, it is instructive to think about how it can fail. It can fail in the first stage by overfitting the endogenous variables or producing predicted values that are so highly correlated that it is hard to tell their effects apart. In addition, NN2SLS can fail in the second stage by failing to include important regressors or including too many unimportant regressors. It is useful to consider these two cases separately. For this purpose, we introduce the oracle NN2SLS estimator. The oracle-NN2SLS estimator is identical to the NN2SLS estimator except that it knows which regressors are "important".

To formally introduce the oracle-NN2SLS estimator, we define the index set

$$S = \{j \in \{1 \dots, p\} | \beta_j^* \neq 0\} \tag{2.14}$$

to be the set of the second-stage coefficients which are "important" i.e. not zero in the data generating process. Denote the cardinality of S by $s := |S|$. We then define the oracle second-stage estimator as

$$\arg\min_{b \in \mathbb{R}_+^s} \frac{1}{n} \left\| y - \hat{X}_S b \right\|_2^2. \tag{oracle-NN2SLS}$$

The difference to (NN2SLS) is that in (oracle-NN2SLS), the set S is known. Hence all β_j's which are zero in the data generating process are also set to zero in (oracle-NN2SLS). Now we can apply

a triangle inequality to split the analyses of first and second stage:

$$\left\|\hat{\beta}^{NN2SLS} - \beta^*\right\|_1 \leqslant \left\|\hat{\beta}^{oracle} - \beta^*\right\|_1 + \left\|\hat{\beta}^{NN2SLS} - \hat{\beta}^{oracle}\right\|_1.$$

Intuitively, the challenge in bounding $\left\|\hat{\beta}^{NN2SLS} - \hat{\beta}^{oracle}\right\|_1$ is the high-dimensional second stage. Both $\hat{\beta}^{NN2SLS}$ and $\hat{\beta}^{oracle}$ are based on the same $\hat{X}$ so that the analysis is separated from the problem of using the d instruments without overfitting. The focus is on avoiding overfitting in the second stage when NNLS selects among p covariates. We control the term $\left\|\hat{\beta}^{NN2SLS} - \hat{\beta}^{oracle}\right\|_1$ using the KKT conditions of the NN2SLS problem and our assumption on instrument relevance formalized in Assumption 5.

In contrast, the term $\left\|\hat{\beta}^{oracle} - \beta^*\right\|_1$ can be thought of as accuracy of a low-dimensional second-stage problem. Since the second stage is low-dimensional, we can focus on the first stage where we have to avoid overfitting. This requires assumptions on the regularizing power of the first stage method, formalized in Assumption 6. The rest of the argument is based on the form of the objective of the NN2SLS oracle estimator, see Lemma 41: because the NN2SLS problem is convex, we can credibly claim that the objective function evaluated at the estimator is not larger than the objective function evaluated at the true parameter. All other lemmas are more or less technical derivations of probabilistic bounds of terms appearing in the proofs of these three results. Our proofs build on the analysis by Meinshausen (2013).

2.4.1.5 Assumptions

The first assumption concerns independence of our observations and exogeneity of the instruments.

Assumption 3. For each i, write $\eta_i = (\eta_{i,1}, \ldots, \eta_{i,p}) \in \mathbb{R}^p$ and $Z_{i,\cdot} = (Z_{i,1}, \ldots, Z_{i,d})$. We assume that

1. the observations are independent over i and
2. for all i, the instruments Z_i are uncorrelated with ε_i and η_i.

The second assumption imposes sub-Gaussian tail conditions for η, ε, X and Z.[11] These tail assumptions are stronger than the usual integrability assumptions and allow a finite-sample analysis without assuming exact distributions, for example Gaussianity of errors.

Assumption 4.

1. There exists a parameter ρ_ε such that for all i and for all $t \geqslant 0$

$$\mathbb{P}\left[|\varepsilon_i| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{\rho_\varepsilon^2}\right).$$

2. There exists parameters ρ_Z, ρ_η such that for any nonzero vectors $a \in \mathbb{R}^d, b \in \mathbb{R}^p$ and for all $t \geqslant 0$

$$\mathbb{P}\left[\left|Z_{i,\cdot}\frac{a}{\|a\|_2}\right| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{\rho_Z^2}\right),$$
$$\mathbb{P}\left[\left|\eta_{i,\cdot}\frac{b}{\|b\|_2}\right| \geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{\rho_\eta^2}\right).$$

In addition to the *instrument exogeneity* assumption in item 2 of Assumption 3, we need to impose an *instrument relevance* assumption. In a low-dimensional 2SLS setting, this is usually done by requiring that $\mathbb{E}[(Z\pi)'(Z\pi)] = \mathbb{E}[\hat{X}'\hat{X}]$ be invertible. When p is larger than n, the matrix $\hat{X}^t\hat{X} \in \mathbb{R}^{p\times p}$ cannot be invertible because its rank is at most n. What would be a sensible rank condition in a high-dimensional case? Studying H2SLS based on applying LASSO as first and second-stage method to estimate a high-dimensional model with endogeneity, Zhu (2018) assumes that the minimum eigenvalue of $\mathbb{E}[\frac{1}{n}\hat{X}^t\hat{X}]$ is bounded away from zero and that $\left|b^t\frac{\hat{X}^t\hat{X}}{n}b\right|$ is, with high probability, bounded from below by a constant times $(\|b\|_1^2 + \|b\|_2^2)$. As we impose non-negativity of β*, we can work with a different assumption on the design matrix proposed by Slawski and Hein (2011) and Meinshausen (2013) (the "self-regularizing property" by Slawski

[11] See section 2.5 of Vershynin (2020) for a textbook exposition.

and Hein (2011) and the "minimum positive eigenvalue condition" by Meinshausen (2013)). This assumption is the first item of Assumption 5. Let us introduce the assumption before discussing it further.

Assumption 5. We assume

1. There exists a $\nu > 0$ such that with probability at least $p_{A3,1,n}$ we have that for all $b \in \mathbb{R}^p_+$

$$b^t \frac{\hat{X}^t \hat{X}}{n} b \geqslant \nu \left\| b \right\|_1^2 . \tag{2.15}$$

2. There exists a $\phi > 0$ such that with probability at least $p_{A3,2,n}$ it holds that for all $b \in \mathbb{R}^p$ such that $\left\| b_N \right\|_1 \leqslant 3\frac{\sqrt{s}}{\sqrt{\nu}} \left(\max_{j \in S} r_{L45}(j) \right) \left\| b_S \right\|_1$ and $\min_{j \in N} b_j \geqslant 0$ that

$$s\, b^t \frac{\hat{X}^t \hat{X}}{n} b \geqslant \phi \left\| b \right\|_1^2 \tag{2.16}$$

   where $r_{L45}(\cdot)$ is defined in (2.44).

3. There exists a $\phi_\infty > 0$ such that with probability 1 it holds for all $b \in \mathbb{R}^s$ that

$$s\, b \frac{\hat{X}_S^t \hat{X}_S}{n} b \geqslant \phi_\infty \left\| b \right\|_1^2 . \tag{2.17}$$

The first item of Assumption 3 is a minimum *positive* eigenvalue condition on $\hat{X}$. The key is to understand that b is restricted to be non-negative in all components. Recall from the discussion above that it is not feasible to assume that $\hat{X}^t \hat{X}$ has full rank, or equivalently, that the smallest eigenvalue in absolute values is not zero. Indeed, when $n \geqslant p$ then

$$\min_{b \in \Diamond} b^t \frac{\hat{X}^t \hat{X}}{n} b = 0,$$

where $\Diamond =: \{w \in \mathbb{R}^p | \left\| w \right\|_1 = 1\}$ is the diamond that is the boundary of the ball of radius 1 around

zero with with respect to the l_1 norm. [12] However, when one requires that the "eigenvectors" b be componentwise non-negative, Slawski and Hein (2011) and Meinshausen (2013) are able to give numerous examples and sufficient conditions for (2.15). Geometrically, (2.15) is a restriction of the b vectors to be in the simplex of $\mathbb{R}^p$ rather than in the complete diamond $\Diamond$.

Now consider the second item of Assumption 5. Compatibility conditions are among the weakest known regularity assumptions in high-dimensional linear models (Candes et al., 2007; van de Geer and Bühlmann, 2009). Meinshausen (2013) imposed a stronger compatibility condition which does not restrict b_N to be non-negative. Because $N = S^c$ contains "almost all" of the p regressors, restricting b_N to be non-negative is a significant weakening of assumptions. Intuitively, while (2.15) requires all components of b to be non-negative, the mostly non-negative compatibility condition (2.16) allows for b_S to be negative in some components but, in exchange, requires that b_N is "not too large" compared to b_S.

Finally, the third item in assumption 5 is a usual invertability condition on $\hat{X}_S^t \hat{X}_S$. It is possible to assume this invertibility because s is much smaller than n, i.e. we assume second-stage sparsity. See Assumption 7 for detailed growth assumptions.

Next, we impose an assumption on the first-stage prediction accuracy that allows us to be flexible with respect to the first-stage method.

Assumption 6. For any subset $\mathfrak{J} \subset \{1, \ldots, p\}$, with probability at least $p_{A4,\mathfrak{J},n}$, we have

$$\frac{1}{\sqrt{n}} \max_{j \in \mathfrak{J}} \left\| Z(\hat{\pi}_j - \pi_j^*) \right\|_2 \leqslant r_{A4,\mathfrak{J},n}. \tag{2.18}$$

Note that Assumption 6 does not require consistency of the first-stage estimates $\hat{\pi}$. In fact, it does not even require uniqueness of the solution to the first-stage problem. Prediction accuracy is a much weaker requirement for which many results are readily available in the literature. One example for a first-stage method [13]which satisfies this requirement is the LASSO. For example,

[12]The normalization is chosen for comparability with 2.15 (divide 2.15 by $\|b\|_1^2$ for non-zero b).

[13]Assumption 6 is formulated for any subset $\mathfrak{J} \subset \{1, \ldots, p\}$. In contrast, results for many methods are available

Corollary 50 in section 2.8.3 in the appendix is a finite-sample result on the so-called slow rate of the LASSO as a first-stage estimator where the only requirement is sufficient regularization. In this case, for any $0 < \tau < \sqrt{\frac{n}{\log(d\|\mathfrak{J}\|)}}$ and

$$\lambda = 2\rho_z\rho_\eta\sqrt{\frac{\log(d\,|\mathfrak{J}|)}{n}}\tau,$$

assumption 6 holds with

$$p_{A4,\mathfrak{J},n}^{\text{LASSO, slow}} := 1 - 2\exp\left(\log(d\,|\mathfrak{J}|)\left(1 - c_B\tau^2\right)\right),$$

where c_B is a mathematical constant and

$$r_{A4,\mathfrak{J},n} := 24\rho_z\rho_\eta\sqrt{\frac{\log(d\,|\mathfrak{J}|)}{n}}\tau\left\|\pi_j^*\right\|_1.$$

In particular, we see that if $\tau > \frac{1}{\sqrt{c_B}}$, then $p_{A4,\mathfrak{J},n}^{\text{LASSO, slow}}$ tends to one if $d\mathfrak{J}$ tends to infinity. We also see that in this case, $r_{A4,\mathfrak{J},n}$ tends to zero at rate $\sqrt{\frac{\log(d|\mathfrak{J}|)}{n}}$.

An applied researcher may have knowledge of the specific structure of the Z matrix. In particular, the Z matrix might satisfy sufficient regularity conditions for some other first-stage estimator. By leaving the first-stage estimator unspecified, we allow the applied researcher to choose a suitable first-stage method depending on the application.

2.4.2 Application to Estimating Nesting Structures

To apply the theory of NN2SLS to the problem of estimating 2.7, we note that n, the number of observations, is the product of J, the number of products, and T, the number of markets. Further, the number of groups or nests is $G = 2^J - J - 1$.

only for $\mathfrak{J}$ with exactly one element. Remark 2.48 in appendix 2.8.2 shows how to use a union bound to go from such results to the one required in Assumption 6.

In the general theory, all second-stage coefficients are non-negative. From economic theory, we know that all nesting parameters are non-negative. It is also reasonable to assume that α, the marginal utility of income, is non-negative.[14] However, we cannot expect that all product characteristics have a non-negative coefficient. We apply Frisch-Waugh-Lovell to residualize the linear model with respect to the product characteristics $\mathbf{X}$. Let

$$M_{\mathbf{X}} := I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

be the matrix projecting on the orthogonal of the columnspace of $\mathbf{X}$. Then we can set

$$X := M_{\mathbf{X}}(p, \mathfrak{G}),$$

where the left-hand side $X \in \mathbb{R}^{n,G+1}$ is the matrix for the general theory, written in non-bold font to distinguish it from the matrix of characteristics, $\mathbf{X} \in \mathbb{R}^{n \times K}$, where we recall that K is the number of product characteristics. Similarly, we set $y := M_{\mathbf{X}}\mathbf{y}$ and $Z := M_{\mathbf{X}}\mathbf{Z}$.

2.5 Monte-Carlo Experiments

We carry out two sets of Monte-Carlo experiments. The first set assesses the performances in finite samples of our proposed estimator (Section 2.5.1). The second set compares our approach to the BLP approach (Section 2.5.2).

[14] This is not necessary but simplifies the notation.

2.5.1 Performances of the Estimator

2.5.1.0.1 Simulated Data.

We generate 50 datasets of $T = 400$ markets of $J = 8$ products (+ one outside good). The simulated data generating process (DGP) is a fully structural model of demand and supply, where the demand model is defined by Equations (2.6), and where the supply model is a static price competition model with multiproduct firms.

On the supply side, we assume that 4 firms, each producing one two products compete in prices: in market t, each firm chooses the prices p_{jt} of its products that maximize its profits. The marginal cost function of product j in market t is given by

$$c_{jt} = \gamma_0 + \gamma_x x_{jt} + \gamma_w w_{jt} + \omega_{jt}, \quad j > 0, \tag{2.19}$$

where w_{jt} is a cost-shifter and ω_{jt} is an unobserved cost component.

Assuming that a pure-strategy Nash equilibrium in prices exists, prices of products $j = 1, \ldots, J$ solve the corresponding first-order conditions, and the associated market shares are then computed using Equations (2.6).

We use x_{jt}, w_{jt} i.i.d. $\mathcal{U}(0,1)$ and $(\xi_{jt}, \omega_{jt}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$. We set $\beta_0 = 3$, $\beta_x = 6$, $\alpha = 1$, $\gamma_0 = 1$, $\gamma_x = 2$, $\gamma_w = 2$. We set $\mu_{\{1,2\}} = \mu_{\{1,3\}} = \mu_{\{2,3\}} = \mu_{\{1,2,3\}} = \mu_{\{1,4\}} = \mu_{\{2,4\}} = 0.1$, where, e.g., $\mu_{\{1,2\}}$ denotes the nesting parameter associated to the group composed of products 1 and 2. The remaining 241 nesting parameters are set equal to zero.

Let $d^x_{ijt} \equiv x_{it} - x_{jt}$. We use the two sets of instruments. The first comprises the cost shifter w_{jt}. The second consists of the differentiation instruments $\left(\sum_{i \in \mathcal{G}(j)} d^x_{ijt}\right)$, $\left(\sum_{i \in \mathcal{G}(j)} \left(d^x_{ijt}\right)^2\right)$ and $\left(\sum_{i \in \mathcal{G}(j)} d^x_{ijt}\right)^2$ that we interact with the corresponding group fixed effects.

2.5.1.0.2 Results.

Table 2.3 presents the results. It shows that our estimator performs well in finite sample. Recall that in the DGP, 6 nesting parameters out of 247 are positive, that is, 2.43% among them are positive (97.57% are zero). Overall, our estimator predicts that 10.36% of the nesting parameters are positive (89.64% are zero). Furthermore, we correctly detect whether the nesting parameters are zero or positive for 91.88% of them. Specifically, we correctly detect as positive (resp., zero) 96% (resp.,91.78%) of the positive (resp., zero) nesting parameters.

Table 2.3: Monte-Carlo Results

	True	Estimates	
β_0	3.000	3.028	(0.0500)
$-\alpha$	-1.000	-0.924	(0.0283)
β_x	6.000	5.576	(0.1293)
Total of			
Zeros	97.57%		89.64%
Non-zeros	2.43%		10.36%
Correctly detected			91.88%
Zeros			91.78%
Non-zeros			96.00%

Notes: Estimates (mean across 50 Monte Carlo datasets); Simulated standard errors into parenthesis (standard deviation across the 50 Monte Carlo datasets).

2.5.2 Comparison to BLP Approach

2.5.2.1 Simulated Data

We generate 50 datasets of $T = 400$ markets of $J = 8$ products (+ one outside good). The simulated DGP is a fully structural model of demand and supply, where the demand model is a random coefficient logit model (RCL) with one log-normally distributed coefficient on price, and where the supply model is a static price competition model with multiproduct firms.

On the demand side, the RCL model is an additive random utility model where the indirect utility of consumer n in market t for product j is defined by

$$u_{njt} = \beta_0 - \alpha p_{jt} + \beta_x x_{jt} - \alpha_n p_{jt} + \xi_{jt} + \varepsilon_{njt}, \quad j > 0 \tag{2.20}$$

$$u_{n0t} = \varepsilon_{n0t} \tag{2.21}$$

where $\alpha_n \sim \log \mathcal{N}(0, \sigma)$ and ε_{njt} i.i.d. $T1EV(0, 1)$. On the supply side, this is the same as in the first experiment, except that market shares are now computed using the RCL demand function.

Simulations (and estimations of the RCL model) use the package `pyblp` in Python, developed by Conlon and Gortmaker (2020). We set $\beta_0 = 5$, $\beta_x = 6$, $\alpha = 2$, $\sigma = 1.5$, $\gamma_0 = 1$, $\gamma_x = \gamma_w = 2$. We use x_{jt}, w_{jt} i.i.d. $\mathcal{U}(0, 1)$ and $(\xi_{jt}, \omega_{jt}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$.

Estimation of the RCL models uses Berry et al. (1995)'s method together with the best practices as described in Conlon and Gortmaker (2020): it first uses Gandhi and Houde (2020)'s differentiation IVs to control for the endogeneity of prices and market shares and then update to the optimal IVs of Chamberlain (1987) using the approximate approach.

Estimation of the IGNL model follows the two steps described above. We use two sets of instruments. Based on Reynaert and Verboven (2014) and Gandhi and Houde (2020), the first comprises the predicted value $\hat{p}_{jt}$ from a linear regression of prices p_{jt} on a third-order polynomial in (x_{jt}, z_{jt}) and a function of ownership and d^x_{ijt}. Let $d^{\hat{p}}_{ijt} = \hat{p}_{it} - \hat{p}_{jt}$. The second set of instruments

consists of the differentiation instruments $\left(\sum_{i\in\mathcal{G}(j)} d_{ijt}^{\hat{p}}\right)$, $\left(\sum_{i\in\mathcal{G}(j)} \left(d_{ijt}^{\hat{p}}\right)^2\right)$ and $\left(\sum_{i\in\mathcal{G}(j)} d_{ijt}^{\hat{p}}\right)^2$ that we interact with the corresponding group fixed effects.

2.5.2.2 Results

The following table summarizes the results of the experiment. It shows that the IGNL model best fits the own- and cross-price elasticities of a RCL model with a log-normally distributed random coefficient on price than a misspecified RCL model in which the random coefficient on prices is assumed to be normal. It also shows that our approach and the BLP approach when it is misspecified perform rather equally in terms of estimated markups.

Table 2.4: Monte Carlo: True Model is RCL with log-normal coefficients

	True	Estimated					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	True	Bias	t-stat	Med. bias	MAD	MSE	Var
True Model: assuming that the random coefficient is log-normal							
Own-elasticities	-7.787	0.009	0.827	0.018	0.045	0.003	0.004
Cross-elasticities	0.770	-0.001	-0.464	-0.002	0.006	0.000	0.000
Markups	0.165	0.000	0.897	0.000	0.001	0.000	0.000
Misspecified Model: assuming that the random coefficient is normal							
Own-elasticities	-7.787	0.624	34.158	0.626	0.624	0.401	0.015
Cross-elasticities	0.770	-0.111	-16.450	-0.100	0.111	0.014	0.002
Markups	0.165	-0.000	-1.3841	-0.001	0.002	0.000	0.000
Misspecified Model: assuming an IGNL model							
Own-elasticities	-7.787	-0.136	-6.837	-0.104	0.141	0.035	0.019
Cross-elasticities	0.770	-0.054	-13.544	-0.058	0.056	0.004	0.001
Markups	0.165	0.013	22.439	0.013	0.013	0.000	0.000

Notes: Results use 50 Monte Carlo datasets. Column (1) gives the true own elasticities, cross elasticities and markups. Column (2) gives the bias, Column (3) gives the t-stat for whether the bias (which is estimated in a finite number of simulations) is statistically different from zero; Column (4) gives median bias, Column (5) gives mean absolute deviation ; Column (6) gives the mean squared error and Column (7) gives the variance.

Table 2.5: Monte Carlo: True Model is RCL with normal coefficients

	True	Estimated					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	True	Bias	t-stat	Med. bias	MAD	MSE	Var
True Model: assuming that the random coefficient is log-normal							
Own-elasticities	-5.262	0.014	1.597	0.017	0.044	0.003	0.003
Cross-elasticities	0.660	-0.002	-1.319	-0.002	0.006	0.000	0.000
Markups	0.234	0.001	1.636	0.001	0.002	0.000	0.000
Misspecified Model: assuming that the random coefficient is normal							
Own-elasticities	-5.262	1.016	34.161	1.039	1.016	1.073	0.044
Cross-elasticities	0.660	-0.194	-96.121	-0.192	0.194	0.039	0.000
Markups	0.234	0.037	1.311	0.062	0.120	0.040	0.039
Misspecified Model: assuming an IGNL model							
Own-elasticities	-5.262	-0.280	-26.426	-0.282	0.280	0.083	0.005
Cross-elasticities	0.660	0.003	1.599	0.003	0.009	0.000	0.000
Markups	0.234	0.008	11.513	0.008	0.008	0.000	0.000

Notes: Results use 50 Monte Carlo datasets. Column (1) gives the true own elasticities, cross elasticities and markups. Column (2) gives the bias, Column (3) gives the t-stat for whether the bias (which is estimated in a finite number of simulations) is statistically different from zero; Column (4) gives median bias, Column (5) gives mean absolute deviation ; Column (6) gives the mean squared error and Column (7) gives the variance.

2.6 Conclusion

We propose a method to estimate the nesting structure in a large class of demand models for differentiated products, which describe the aggregate behavior of heterogeneous, utility-maximizing

consumers.

We build on a recent generalization of the inverse demand function of the nested logit model that accommodates any possible nesting structure while retaining its tractability. Specifically, we show how using non-negativity restrictions coming from economic theory as well as sparsity restrictions help us estimate the nesting structure.

2.7 Bibliography

Aboutaleb, Y. M., M. Danaf, Y. Xie, and M. Ben-Akiva (2021). Discrete choice analysis with machine learning capabilities. *arXiv preprint arXiv:2101.10261*.

Allen, R. and J. Rehbeck (2019). Identification with additively separable heterogeneity. *Econometrica 87*(3), 1021–1054.

Almagro, M. and E. Manresa (2019). Data-driven nests in discrete choice models. *Available on authors' websites*.

Azar, J., S. Berry, and I. E. Marinescu (2019). Estimating labor market power. *Available at SSRN 3456277*.

Belloni, A., V. Chernozhukov, C. Hansen, and W. Newey (2017). Simultaneous confidence intervals for high-dimensional linear models with many endogenous variables. *arXiv preprint arXiv:1712.08102*.

Berry, S., A. Eizenberg, and J. Waldfogel (2016). Optimal product variety in radio markets. *The RAND Journal of Economics 47*(3), 463–497.

Berry, S. and P. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5), 1749–1797.

Berry, S. and P. Haile (2016). Identification in differentiated products markets. *Annual review of Economics 8*, 27–52.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*, 841–890.

Berry, S., J. Levinsohn, and A. Pakes (1999). Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review 89*(3), 400–430.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics 25*, 242–262.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics 37*(4), 1705–1732.

Björnerstedt, J. and F. Verboven (2016). Does merger simulation work? evidence from the swedish analgesics market. *American Economic Journal: Applied Economics 8*(3), 125–64.

Boyd, J. H. and R. E. Mellman (1980). The effect of fuel economy standards on the us automotive market: an hedonic demand analysis. *Transportation Research Part A: General 14*(5-6), 367–378.

Brenkers, R. and F. Verboven (2006). Liberalizing a distribution system: the european car market. *Journal of the European Economic Association 4*(1), 216–251.

Bresnahan, T. F., S. Stern, and M. Trajtenberg (1997). Market segmentation and the sources of rents from innovation: Personal computers in the late 1980s. *RAND Journal of Economics*, S17–S44.

Candes, E., T. Tao, et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of statistics 35*(6), 2313–2351.

Cardell, N. S. and F. C. Dunbar (1980). Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General 14*(5-6), 423–434.

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

Compiani, G. (2020). Market counterfactuals and the specification of multi-product demand: A nonparametric approach. *Unpublished, available on author's website*.

Conlon, C. and J. Gortmaker (2020). Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics 51*(4), 1108–1161.

Davis, P. and P. Schiraldi (2014). The flexible coefficient multinomial logit (fc-mnl) model of demand for differentiated products. *The RAND Journal of Economics 45*(1), 32–63.

Dubé, J.-P. (2018). Microeconometric models of consumer demand. *Booth School of Business, University of Chicago*.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica 80*(5), 2231–2267.

Fosgerau, M., J. Monardo, and A. de Palma (2021). The inverse product differentiation logit model. *Available at SSRN: https://ssrn.com/abstract=3141041*.

Gandhi, A. and J.-F. Houde (2020). Measuring substitution patterns in differentiated products industries. *Available at SSRN: https://ssrn.com/abstract=3472810*.

Gautier, E. and A. B. Tsybakov (2018). High-dimensional instrumental variables regression and confidence sets–v2/2012. *arXiv preprint arXiv:1812.11330*.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *The American Economic Review 97*(3), 713–744.

Goldberg, P. K. (1995). Product differentiation and oligopoly in international markets: The case of the us automobile industry. *Econometrica*, 891–951.

Griffith, R., M. O'Connell, and K. Smith (2019). Tax design in the alcohol market. *Journal of Public Economics 172*, 20–35.

Grigolon, L. (2020). Blurred boundaries: a flexible approach for segmentation applied to the car market. *Unpublished, available on author's website*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hausman, J., G. Leonard, and J. D. Zona (1994). Competitive analysis with differenciated products. *Annales d'Economie et de Statistique*, 159–180.

Horowitz, J. (1987). Specification tests for nested logit models. *Environment and Planning A 19*(3), 395–402.

Knittel, C. R. and K. Metaxoglou (2014). Estimation of random-coefficient demand models: Two empiricists' perspective. *Review of Economics and Statistics 96*(1), 34–59.

Koppelman, F. S. and C.-H. Wen (2000). The paired combinatorial logit model: properties, estimation and application. *Transportation Research Part B: Methodological 34*(2), 75–89.

Lee, J. and K. Seo (2015). A computationally fast estimator for random coefficients logit demand models using aggregate data. *The RAND Journal of Economics 46*(1), 86–102.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record* (673).

McFadden, D. and K. Train (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics 15*(5), 447–470.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics 7*, 1607–1631.

Miller, N. H. and M. C. Weinberg (2017). Understanding the price effects of the millercoors joint venture. *Econometrica 85*(6), 1763–1791.

Nevo, A. (2000). Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The RAND Journal of Economics 31*(3), 395–421.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica 69*(2), 307–342.

Nevo, A. (2011). Empirical models of consumer behavior. *Annu. Rev. Econ. 3*(1), 51–75.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy 110*(4), 705–729.

Reynaert, M. and F. Verboven (2014). Improving the performance of random coefficients demand models: the role of optimal instruments. *Journal of Econometrics 179*(1), 83–98.

Salanié, B. and F. A. Wolak (2019). Fast," robust", and approximately correct: estimating mixed demand systems. *National Bureau of Economic Research*.

Slawski, M. and M. Hein (2011). Sparse recovery by thresholded non-negative least squares. *Advances in neural information processing systems 24*.

Slawski, M. and M. Hein (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics 7*, 3004–3056.

Small, K. A. (1987). A discrete choice model for ordered alternatives. *Econometrica*, 409–424.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist. 3*, 1360–1392.

Verboven, F. (1996a). International price discrimination in the european car market. *The RAND Journal of Economics*, 240–268.

Verboven, F. (1996b). The nested logit model and representative consumer theory. *Economics Letters 50*(1), 57–63.

Vershynin, R. (February 13, 2020). High-dimensional probability: An introduction with applications in data science. *URL: http://www-personal. umich. edu/˜ romanv/papers/HDP-book/HDP-book. pdf*.

Wen, C.-H. and F. S. Koppelman (2001). The generalized nested logit model. *Transportation Research Part B: Methodological 35*(7), 627–641.

Zhu, Y. (2018). Sparse linear models and l1-regularized 2sls with high-dimensional endogenous regressors and instruments. *Journal of Econometrics 202*(2), 196–213.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

2.8 Appendix

2.8.1 Non-negative Two Stage Least Squares

Let us first define some important constants. By c_H we denote the constant that appears in Hoeffding's inequality for sub-Gaussians ((Vershynin, 2020, Theorem 2.6.3)).[15] By $c_{center,SG}$, we denote the constant that appears in the centering inequality for sub-Gaussians, see (Vershynin, 2020, Lemma 2.6.8). Similarly, by c_B we denote the constant that appears in Bernstein's inequality for sub-exponential random variables (Theorem 2.6.3 in Vershynin (2020)).[16] By $c_{center,SE}$, we denote the constant that appears in the centering inequality for sub-exponentials, see (Vershynin, 2020, Exercise 2.7.10).

2.8.1.1 Proof of Theorem 38.

Proof. In general, the proof follows the arguments in Meinshausen (2013). Using the triangle inequality, we see

$$\begin{aligned}\left\|\hat{\beta}^{NN2SLS}-\beta^{*}\right\|_1 &= \left\|\hat{\beta}^{oracle}-\beta^{*}-\left(\hat{\beta}^{oracle}-\hat{\beta}^{NN2SLS}\right)\right\|_1 \\ &\leqslant \left\|\hat{\beta}^{oracle}-\beta^{*}\right\|_1+\left\|\hat{\beta}^{NN2SLS}-\hat{\beta}^{oracle}\right\|_1 .\end{aligned}$$

The term $\left\|\hat{\beta}^{oracle}-\beta^{*}\right\|_1$ can be bounded using Lemma 41.

[15] Note that this constant also depends on which of the "equivalent" definitions of a sub-Gaussian one considers since they only equivalent up to a constant, (Vershynin, 2020, Proposition 2.5.2).

[16] Note that this constant also depends on which of the "equivalent" definitions of a sub-exponential one considers since they only equivalent up to a constant, (Vershynin, 2020, Proposition 2.7.1).

Upper bound for $\Delta\beta^t \frac{\hat{X}^t\hat{X}}{n}\Delta\beta$

To estimate the second term, define $\Delta\beta := \hat{\beta}^{NN2SLS} - \hat{\beta}^{oracle}$.[17] We can reparametrize the NN2SLS problem (2.10) problem so that the solution is $\Delta\beta$:

$$\Delta\beta = \min_{\gamma} \underbrace{\left\| Y - \hat{X}\hat{\beta}^{oracle} - \hat{X}\gamma \right\|_2^2}_{=:f(\gamma)} \qquad \text{s.t } \gamma_k \geqslant -\hat{\beta}_k^{\text{oracle}} \qquad \text{for all } k = 1, \ldots, p. \tag{2.22}$$

Note that $\gamma = 0$ is a feasible point to (2.22). Hence $f(0) \geqslant f(\Delta\beta)$, i.e.

$$\left\| Y - \hat{X}\hat{\beta}^{oracle} \right\|_2^2 \geqslant \left\| Y - \hat{X}\hat{\beta}^{oracle} - \hat{X}\Delta\beta \right\|_2^2 \tag{2.23}$$

Defining $R := Y - \hat{X}\hat{\beta}^{oracle}$. Then we can write inequality (2.23) as

$$R^t R \geqslant \left\| R - \hat{X}\Delta\beta \right\| = R^t R - 2R^t\hat{X}\Delta\beta + \Delta\beta^t\hat{X}^t\hat{X}\Delta\beta.$$

Subtracting the term R^tR, dividing by n and rearranging yields

$$\Delta\beta^t \frac{\hat{X}^t\hat{X}}{n}\Delta\beta \leqslant \frac{2}{n}R^t\hat{X}\Delta\beta.$$

Using Lemma (40), we see that with probability at least p_{L2} we have

$$\Delta\beta^t \frac{\hat{X}^t\hat{X}}{n}\Delta\beta \leqslant r_{L2} \left\| \Delta\beta_N^* \right\|_1 \leqslant r_{L2} \left\| \Delta\beta_{M^c} \right\|_1. \tag{2.24}$$

[17] Meinshausen writes $\delta\beta$ instead of $\Delta\beta$.

Lower bound for $\Delta\beta^t \frac{\hat{X}^t\hat{X}}{n}\Delta\beta$

Define $M := \{k : \Delta\beta_k < 0\}$. By definition[18] [19], $M \subset S$ and $N \subset M^c$ so that we trivially have $\|\Delta\beta_N\|_1 \leqslant \|\Delta\beta_{M^c}\|_1$. Now define

$$a := \sqrt{\Delta\beta^t_{M^c}\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_{M^c}},$$
$$b := \sqrt{\Delta\beta^t_{M}\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_{M}}$$

where, *exceptionally*, we write $\Delta\beta_{M^c}$ to for the p-dimensional vector defined by

$$(\Delta\beta_{M^c})_j = \begin{cases} \Delta\beta_j & \text{if } j \in M^c, \\ 0 & \text{otherwise} \end{cases}$$

and similarly for $\Delta\beta^t_M$. Then we have

$$\begin{aligned}
&\Delta\beta^t\frac{\hat{X}^t\hat{X}}{n}\Delta\beta \\
=&\left(\Delta\beta^t_M + \Delta\beta^t_{M^c}\right)^t\frac{\hat{X}^t\hat{X}}{n}\left(\Delta\beta^t_M + \Delta\beta^t_{M^c}\right) \\
=&\Delta\beta^t_M\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_M + \Delta\beta^t_{M^c}\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_{M^c} + 2\Delta\beta^t_{M^c}\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_M \\
=&a^2 + b^2 + 2\Delta\beta^t_{M^c}\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_M \\
\geqslant&a^2 + b^2 - 2\left|\Delta\beta^t_{M^c}\frac{\hat{X}^t\hat{X}}{n}\Delta\beta_M\right| \\
\geqslant&a^2 + b^2 - 2ab && \text{(Cauchy-Schwarz)} \\
\geqslant&a^2 - 2ab. && (b \geqslant 0)
\end{aligned}$$

[18] If $j \in M$, then $\Delta\beta_j = \hat{\beta}^{\text{NN2SLS}}_j - \hat{\beta}^{\text{oracle}}_j < 0$. Because $\hat{\beta}^{\text{NN2SLS}}_j \geqslant 0$, this implies $\hat{\beta}^{\text{oracle}} > 0$ which in turn implies that $j \in S$.

[19] If $j \in N$, then $\hat{\beta}^{\text{oracle}} = 0$ so that $\Delta\beta_j = \hat{\beta}^{\text{NN2SLS}}_j - \hat{\beta}^{\text{oracle}}_j = \hat{\beta}^{\text{NN2SLS}}_j \geqslant 0$ which means that $j \in M^c$.

$$
\begin{aligned}
&\geqslant \nu \left\|\Delta\beta_{M^c}\right\|_1^2 - 2\sqrt{\nu}\left\|\Delta\beta_{M^c}\right\|_1 b && \text{(Pos. Eigenval. Cond.)}\\
&\geqslant \nu \left\|\Delta\beta_{M^c}\right\|_1^2 - 2\sqrt{\nu}\left\|\Delta\beta_{M^c}\right\|_1 \sqrt{\sigma_{\max}\left(\frac{\hat{X}_S^t\hat{X}_S}{n}\right)}\left\|\Delta\beta_M\right\|_2\\
&\geqslant \nu \left\|\Delta\beta_{M^c}\right\|_1^2 - 2\sqrt{\nu}\left\|\Delta\beta_{M^c}\right\|_1 \sqrt{\sigma_{\max}\left(\frac{\hat{X}_S^t\hat{X}_S}{n}\right)}\left\|\Delta\beta_M\right\|_1\\
&\geqslant \nu \left\|\Delta\beta_{M^c}\right\|_1^2 - 2\sqrt{\nu}\left\|\Delta\beta_{M^c}\right\|_1 \sqrt{s}\left(\max_{j\in S} r_{L45}(j)\right)\left\|\Delta\beta_M\right\|_1 && \text{(Lemma 44)}
\end{aligned}
$$

with probability at least $p_{L45}(S) + p_{A3,1,n} - 1$.

Case 1: $\left\|\Delta\beta_M\right\|_1 < \frac{\sqrt{\nu}}{3\sqrt{s}\left(\max\limits_{j\in S} r_{L45}(j)\right)}\left\|\Delta\beta_{M^c}\right\|_1$

In this case we have with probability at least $p_{L45}(S) + p_{A3,1,n} - 1$

$$
\Delta\beta^t\frac{\hat{X}^t\hat{X}}{n}\Delta\beta \geqslant \left\|\Delta\beta_{M^c}\right\|_1^2\nu\left(1-\frac{2}{3}\right) = \frac{1}{3}\left\|\Delta\beta_{M^c}\right\|_1^2\nu.
$$

Combining this with the upper bound (2.24), we find that with probability at least $p_{L2} + p_{A3,1,n} - 1$

$$
\frac{1}{3}\left\|\Delta\beta_{M^c}\right\|_1^2\nu \leqslant r_{L2}\left\|\Delta\beta_{M^c}\right\|_1.
$$

Dividing by $\left\|\Delta\beta_{M^c}\right\|_1$, we find that with probability at least $p_{L2} + p_{A3,1,n} - 1$

$$
\left\|\Delta\beta_{M^c}\right\|_1 \leqslant \frac{3}{\nu}r_{L2}.
$$

Combining this with the assumptions of case 1, we also get a bound for $\left\|\Delta\beta_M\right\|_1$:

$$
\left\|\Delta\beta\right\|_1 = \left\|\Delta\beta_{M^c}\right\|_1 + \left\|\Delta\beta_M\right\|_1
$$

$$\leqslant \|\Delta\beta_{M^c}\|_1 \left(1 + \frac{\sqrt{\nu}}{3\sqrt{s}\left(\max_{j\in S} r_{L45}(j)\right)}\right)$$
$$\leqslant r_{L2}\left(\frac{3}{\nu} + \frac{1}{\sqrt{\nu}\sqrt{s}\left(\max_{j\in S} r_{L45}(j)\right)}\right).$$

Case 2: $\|\Delta\beta_M\|_1 \geqslant \frac{\sqrt{\nu}}{3\sqrt{s}\left(\max_{j\in S} r_{L45}(j)\right)} \|\Delta\beta_{M^c}\|_1$

First, recall that $N \subset M^c$ and $M \subset S$. Hence the assumption for this case implies

$$\|\Delta\beta_N\|_1 \frac{\sqrt{\nu}}{3\sqrt{s}\left(\max_{j\in S} r_{L45}(j)\right)} \leqslant \|\Delta\beta_{M^c}\|_1 \frac{\sqrt{\nu}}{3\sqrt{s}\left(\max_{j\in S} r_{L45}(j)\right)} \leqslant \|\Delta\beta_M\|_1 \leqslant \|\Delta\beta_S\|_1 .$$

In particular, $\Delta\beta$ satisfies the condition required for (2.16) of Assumption 5. Using (2.16), we directly get with probability at least $p_{A3,2,n}$

$$\Delta\beta^t \frac{\hat{X}^t\hat{X}}{n}\Delta\beta \geqslant \frac{\phi}{s}\|\Delta\beta\|_1^2 .$$

Combining this with the upper bound (2.24), we find that with probability at least $p_{L2} + p_{A3,2,n} - 1$

$$\frac{\phi}{s}\|\Delta\beta\|_1^2 \leqslant r_{L2}\|\Delta\beta_{M^c}\|_1 \leqslant r_{L2}\|\Delta\beta\|_1 .$$

Dividing by $\|\Delta\beta\|_1$, we find that with probability at least $p_{L2} + p_{A3,2,n} - 1$

$$\|\Delta\beta\|_1 \leqslant \frac{s}{\phi} r_{L2}.$$

Hence we have bounded $\|\Delta\beta\|_1$ in both cases and completed the proof. □

2.8.1.2 Proof of Corollary 39.

Before we go to the proof of Corollary 39, we formulate an assumptions on growth rates.

Assumption 7. We have

1. $p_{A3,1,n}$ and $p_{A3,2,n}$ converge to 1 as n tends to infinity
2. for any subset $\mathfrak{J} \subset \{1, \ldots, p\}$, $r_{A4,\mathfrak{J},n}$ converges to 0 and $p_{A4,\mathfrak{J},n}$ converges to 1 as n tends to infinity
3. $s^3 r_{A4,S,n}$ converges to zero as n tends to infinity
4. $\frac{\log(d)}{n^{1-2\gamma}}$ converges to zero as n tends to infinity
5. $\frac{s^{\frac{5}{2}}}{n^{\gamma}}$ converges to zero as n tends to infinity
6. $\frac{\log(p)}{n^{1-2\alpha}}$ converges to zero as n tends to infinity
7. $\frac{s^{\frac{3}{2}}}{n^{\alpha}}$ converges to zero as n tends to infinity
8. there exists a universal constant $\overline{\beta}$, independent of n, such that $\left\|\beta^*\right\|_2 \leqslant \sqrt{s}\overline{b}$.
9. there exist a universal constant $\overline{\pi}$, independent of n, such that

$$\max_{j \in \{1,\ldots,p\}} \left\|\pi_j^*\right\|_1 \leqslant \overline{\pi}$$

Proof. To derive consistency from the finite sample result in Theorem 38, we have to do two things. First, we have to show that p_{T1} converges to 1 as n tends to infinity. Secondly, we have to show that r_{T1} converges to zero as n tends to infinity. So let's proceed in two steps.

Step 1: $\mathbf{p_{T1}}$ converges to 1

Recall from Theorem 38 that

$$p_{T1} := p_{L2} + p_{A3,1,n} + p_{A3,2,n} - 2.$$

$\mathbf{p_{L2}}$ converges to 1 as n tends to infinity Recall from Lemma 40 that

$$p_{L2} := p_{L3} + p_{L4} + p_{L45}(\{1, \ldots, p\}) - 2.$$

To show that p_{L2} converges to 1 as n tends to infinity, it is sufficient to show that p_{L3}, p_{L4} and $p_{L45}(\{1, \ldots, p\})$ converge to 1 as n tends to infinity. Consider first p_{L3}. We recall

$$p_{L3} := p_{A4,S,n} + p_{L45}(S) - 1 - 4\exp(-c_B n) - 4\exp(\log(d) - c_B n^{1-2\gamma}).$$

As before, $p_{A4,S,n}$ converges to 1 as n tends to infinity by Assumption 7. In Assumption 7, we have also assumed that $\frac{\log(d)}{n^{1-2\gamma}}$ converges to zero as n tends to infinity so that $\exp(\log(d) - c_B n^{1-2\gamma})$ converges to zero as n tends to infinity. If n tends to infinity, then $\exp(-c_B n)$ trivially converges to zero. Finally, let's consider $p_{L45}(S)$. We recall

$$p_{L45}(S) = p_{A4,S,n} - 2\sum_{j \in S} \exp\left(-c_B \min\{\kappa_j, \kappa_j^2\} n\right).$$

As we have already discussed, $p_{A4,S,n}$ converges to 1 as n tends to infinity by Assumption 7. So we only have to show that $\sum_{j \in S} \exp\left(-c_B \min\{\kappa_j, \kappa_j^2\} n\right)$ converges to zero as n tends to infinity. We have

$$\sum_{j \in S} \exp\left(-c_B \min\{\kappa_j, \kappa_j^2\} n\right) \leqslant s \max_{j \in S} \exp\left(-c_B \min\{\kappa_j, \kappa_j^2\} n\right)$$

$$\leqslant s \exp\left(-c_B \min_{j\in S}\left(\min\{\kappa_j, \kappa_j^2\}\right) n\right)$$
$$= \exp\left(\log(s) - c_B \min_{j\in S}\left(\min\{\kappa_j, \kappa_j^2\}\right) n\right).$$

In Assumption 7, we assumed that $\min_{j\in S} \kappa_j$ is bounded away from zero uniformly for all n and that $\frac{\log(s)}{n}$ converges to zero. Hence $\sum_{j\in S} \exp\left(-c_B \min\{\kappa_j, \kappa_j^2\} n\right)$ converges to zero as n tends to infinity. So we conclude that p_{L3} indeed converges to 1 as n tends to infinity.

Now let's consider p_{L4}. We recall

$$\begin{aligned}
& p_{L4} \\
=& p_{S2} + p_{S3} + p_{S4}(\alpha) - 2 \\
=& \left(p_{A4,S,n} + p_{L45}(\{1,\ldots,p\}) - 1\right) \\
& + \left(p_{A4,\{1,\ldots,p\},n} - 4\exp(-c_B n)\right) \\
& + \left(1 - 4\exp\left(\log(p) - c_B n^{1-2\alpha}\right)\right) - 2.
\end{aligned}$$

We recall that by Assumption 7, $p_{A4,S,n}$ and $p_{A4,\{1,\ldots,p\},n}$ converge to one as n tends to infinity. It is also clear that $\exp(-c_B n)$ converges to zero as n tends to infinity. By Assumption 7, we also have that $\frac{\log(p)}{n^{1-2\alpha}}$ converges to zero as n tends to infinity. Hence $\exp\left(\log(p) - c_B n^{1-2\alpha}\right)$ converges to zero as n tends to infinity. So to see that p_{L4} converges to 1 as n tends to infinity, it only remains to show that $p_{L45}(\{1,\ldots,p\})$ converges to 1 as n tends to infinity. For this, we recall that

$$p_{L45}(\{1,\ldots,p\}) = p_{A4,\{1,\ldots,p\},n} - 2p\exp(-c_B n).$$

By Assumption 7, $p_{A4,\{1,\ldots,p\},n}$ converges to 1 as n tends to infinty. We have also assumed that $\frac{\log(p)}{n}$ converges to zero as n tends to infinity so that $p\exp(-c_B n)$ converges to zero as n tends to infinity. Hence we see that $p_{L45}(\{1,\ldots,p\})$ converges to 1 as n tends to infinity. This completes

the proof that p_{L2} converges to 1 as n tends to infinity.

Step 2: $\mathbf{r_{T1}}$ converges to 0

Recall that

$$r_{T1} = r_{L3} + r_{L2} \max \left\{ \frac{s}{\phi}, \frac{3}{\nu} + \frac{1}{\sqrt{\nu}\sqrt{s}\left(\max_{j \in S} r_{L45}(j)\right)} \right\}.$$

It is sufficient to show that r_{L3} and sr_{L2} converge to zero as n tends to infinity.

r_{L3} **converges to zero** Recall that

$$r_{L3} = 2s\sqrt{\frac{s}{\phi_\infty}} \left(\left\|\beta^*\right\|_\infty r_{A4,S,n} + \frac{\rho_\varepsilon + \rho_\eta \left\|\beta^*\right\|_2}{\min_{j \in S} r_{L45}(j)} \left(\frac{\rho_Z}{n^\gamma} \max_{j \in S} \left\|\pi_j^*\right\|_1 + \sqrt{2c_{center,SE}} r_{A4,S,n} \right) \right).$$

That the terms $s^{\frac{3}{2}} r_{A4,S,n}$, $\frac{s^{\frac{5}{2}}}{n^\gamma}$ and $s^{\frac{5}{2}} r_{A4,S,n}$ converge to zero follows immediately from assumption 7. This implies that r_{L3} converges to zero as n tends to infinity.

$sr_{L4}(\alpha)$ **converges to zero** Recall that

$$sr_{L4}(\alpha) = sr_{S2} + sr_{S3} + sr_{S4}(\alpha).$$

Now let's consider each of these terms separately. First, for sr_{S2} we have

$$sr_{S2} = s^2 \left\|\beta^*\right\|_\infty r_{A4,S,n} \underbrace{\max_{k \in \{1,\ldots,p\}} r_{L5}(k)}_{\text{bounded under assumption 7}}$$

which converges to zero as $s^{\frac{5}{2}} r_{A4,S,n}$ converges to zero by Assumption 7. Secondly, for r_{S3}, we have

$$\begin{aligned} sr_{S3} =& 2s\left(\rho_\eta + \left\|\beta^*\right\|_2\right) r_{A4,\{1,\ldots,p\},n} \\ \leqslant& 2s\left(\rho_\eta^2 + \sqrt{s\beta}^2\right) r_{A4,\{1,\ldots,p\},n} \end{aligned}$$

which also converges to zero because we assumed that $s^{\frac{3}{2}} r_{A4,\{1,\ldots,p\},n}$ converges to zero in Assumption 7. Finally,

$$\begin{aligned} sr_{S4}(\alpha) =& \; s\rho_z\left(\rho_\eta \left\|\beta^*\right\|_2 + \rho_\varepsilon\right) \frac{1}{n^\alpha} \max_{j\in\{1,\ldots,p\}} \left\|\pi_j^*\right\|_2 \\ \leqslant& \; s\rho_z\left(\rho_\eta\sqrt{s\beta} + \rho_\varepsilon\right) \frac{1}{n^\alpha}\overline{\pi} \end{aligned}$$

converges to zero as n tends to infinity as $\frac{s^{\frac{3}{2}}}{n^\alpha}$ converges to zero as n tends to infinity by Assumption 7. This completes the proof. □

2.8.1.3 Technical Lemmata

Lemma 40. Suppose Assumptions 3, 4, 5 and 6 are satisfied. Define $\Delta\beta := \hat{\beta}^{\text{NN2SLS}} - \hat{\beta}^{oracle}$ and $R := Y - \hat{X}\hat{\beta}^{oracle}$. Then with probability at least

$$p_{L2} := p_{L3} + p_{L4} + p_{L45}(\{1,\ldots,p\}) - 2$$

it holds that

$$\frac{1}{n} R^t \hat{X}\Delta\beta \leqslant r_{L2}\left\|\Delta\beta_N\right\|_1,$$

where

$$r_{L2} = r_{L3} \max_{j \in \{1,\dots,p\}} \left(r_{L45}(j)\right)^2 + r_{L4}$$

and p_{L3}, r_{L3} are defined in Lemma 41, p_{L4} and r_{L4} are defined in Lemma 43, and p_{L4} and r_{L4} are defined in Lemma 45, respectively.

Proof. First, we can write

$$\frac{1}{n} R^t \hat{X} \Delta\beta = \frac{1}{n} \sum_{k \in S} \left(R^t \hat{X}_k\right) \Delta\beta_k + \frac{1}{n} \sum_{k \in N} \left(R^t \hat{X}_k\right) \Delta\beta_k. \tag{2.25}$$

Consider first the sum over $k \in S$. Recall that $\hat{\beta}^{oracle}$ is the solution to (oracle-NN2SLS). In particular, it satisfies the KKT conditions[20] for $k \in S$

1. $\hat{\beta}_k^{oracle} > 0$ and $R^t \hat{X}_k = 0$ or
2. $\hat{\beta}_k^{oracle} = 0$ and $R^t \hat{X}_k \leqslant 0$.

The contribution of all cases in (1.) vanishes in (2.25) because $R^t \hat{X}_k = 0$. For $k \in S$ that fall into category (2.), it follows by the non-negativity of $\hat{\beta}_k^{\text{NN2SLS}}$ and $\hat{\beta}_k^{oracle} = 0$ that $\Delta\beta_k \geqslant 0$ and hence

[20] The Lagrangian for this problem is

$$\mathcal{L}(b, \lambda) = \left\| y - \hat{X}_S b \right\|_2^2 - \sum_{j \in S} \lambda_j b_j$$

where $\lambda_j \geqslant 0$ are Lagrange multipliers. Then the KKT conditions are that $\hat{\beta}^{oracle}$ for all $k \in S$

$$0 = 2\hat{X}_k^t \left(y - \hat{X}_S \hat{\beta}^{oracle}\right) + \lambda_k = 2\hat{X}_k^t R + \lambda_k, \qquad \text{(first order condition)}$$

$$0 = \lambda_k \hat{\beta}_k^{oracle}. \qquad \text{(complementary slackness condition)}$$

In particular, if $\hat{\beta}_k^{oracle} > 0$, then $\lambda_k = 0$ so that the first order condition simplifies to $R^t \hat{X}_k = 0$. If $\hat{\beta}_k^{oracle} = 0$, then we cannot tell whether λ_k is equal to zero or not. But in any case, we know that $\lambda_k \geqslant 0$ so that the first order condition implies $\hat{X}_k^t R \leqslant 0$.

$\left(R^t\hat{X}_k\right)\Delta\beta_k \leqslant 0$. We are left with contributions from $k \in N$ in (2.25):

$$\frac{1}{n}R^t\hat{X}\Delta\beta \leqslant \frac{1}{n}\sum_{k\in N}\left(R^t\hat{X}_k\right)\Delta\beta_k \leqslant \frac{1}{n}\max_{k\in N}\left(R^t\hat{X}_k\right)\left\|\Delta\beta_N\right\|_1.$$

It remains to bound $\frac{1}{n}\max_{k\in N}\left(R^t\hat{X}_k\right)$. For this, we write, as in Zhu

$$R = Y - \hat{X}\hat{\beta}^{oracle} = \hat{X}\left(\beta^* - \hat{\beta}^{oracle}\right) + \xi,$$

where

$$\xi := \left(X^* - \hat{X}\right)\beta^* + \eta\beta^* + \varepsilon.$$

Then

$$\begin{aligned}
\frac{1}{n}\max_{k\in N}\left(R^t\hat{X}_k\right) &= \frac{1}{n}\max_{k\in N}\left[\left(\hat{X}_S\left(\beta^*_S - \hat{\beta}^{oracle}\right) + \xi\right)^t\hat{X}_k\right] \\
&= \frac{1}{n}\max_{k\in N}\left[\left(\beta^*_S - \hat{\beta}^{oracle}\right)^t\hat{X}^t_S\hat{X}_k + \xi^t\hat{X}_k\right] \\
&\leqslant \frac{1}{n}\max_{k\in N}\left|\left(\beta^*_S - \hat{\beta}^{oracle}\right)^t\hat{X}^t_S\hat{X}_k\right| + \frac{1}{n}\max_{k\in N}\xi^t\hat{X}_k \\
&\leqslant \left\|\beta^*_S - \hat{\beta}^{oracle}_S\right\|_1\frac{1}{n}\max_{k\in N}\left\|\hat{X}^t_S\hat{X}_k\right\|_\infty + \frac{1}{n}\max_{k\in\{1,\ldots,p\}}\xi^t\hat{X}_k \\
&\leqslant \left\|\beta^* - \hat{\beta}^{oracle}\right\|_1\frac{1}{n}\max_{k\in N}\left\|\hat{X}^t_S\hat{X}_k\right\|_\infty + \frac{1}{n}\left\|\xi^t\hat{X}\right\|_\infty.
\end{aligned}$$

We can use Lemma 41 to see bound $\left\|\beta^* - \hat{\beta}^{oracle}\right\|_1$ and Lemma 43 to bound $\frac{1}{n}\left\|\xi^t\hat{X}\right\|_\infty$. It remains is to bound the term $\frac{1}{n}\max_{k\in N}\left\|\hat{X}_S\hat{X}_k\right\|_\infty$. [21] We have

$$\frac{1}{n}\max_{k\in N}\left\|\hat{X}_S\hat{X}_k\right\|_\infty = \frac{1}{n}\max_{k\in N}\max_{j\in S}\left|\hat{X}'_j\hat{X}_k\right|$$

[21] Conditional on data, we can compute this quantity, but as we see in the following, we can get an unconditional bound under our assumptions.

$$
\begin{aligned}
&\leqslant \frac{1}{n} \max_{k \in N} \max_{j \in S} \left\| \hat{X}_j \right\|_2 \left\| \hat{X}_k \right\|_2 && \text{(Hölder)} \\
&= \left(\frac{1}{\sqrt{n}} \max_{j \in S} \left\| \hat{X}_j \right\|_2 \right) \left(\max_{k \in N} \frac{1}{\sqrt{n}} \left\| \hat{X}_k \right\|_2 \right).
\end{aligned}
$$

Now we can apply Lemma 45 to bound these two terms. We conclude by combining these bounds with a union bound. □

Lemma 41. Fix $\gamma \in \left(0, \frac{1}{2}\right)$. Under Assumption 3, 4, 5 and 6 with probability at least

$$
p_{L3} := p_{A4,S,n} + p_{L45}(S) - 1 - 4\exp(-c_B n) - 4d\exp(-c_B n^{1-2\gamma})
$$

it holds that

$$
\left\| \hat{\beta}^{oracle} - \beta^* \right\|_1 \leqslant r_{L3}
$$

where

$$
r_{L3} = 2s\sqrt{\frac{s}{\phi_\infty}} \left(\left\| \beta^* \right\|_\infty r_{A4,S,n} + \frac{\rho_\varepsilon + \rho_\eta \left\| \beta^* \right\|_2}{\min_{j \in S} r_{L45}(j)} \left(\frac{\rho_Z}{n^\gamma} \max_{j \in S} \left\| \pi_j^* \right\|_1 + \sqrt{2c_{center,SE}} r_{A4,S,n} \right) \right)
$$

and $p_{L45}(\cdot)$ and $r_{L45}(\cdot)$ are defined in Lemma 45.

Remark 42. As Lemma 41 considers the oracle problem in the second stage, all assumptions concerning first stages for x_j with $j \in S^c$ are not needed. In particular, we do not require

- sub-Gaussianity of $Z_{i,\cdot}$ and $\eta_{i,\cdot}$ of Assumption 4 but only sub-Gaussianity $Z_{i,S}$ and $\eta_{i,S}$
- condition (2.15) and (2.16) of Assumption 5
- Assumption 6 with $\mathfrak{J} \cap S^c \neq \varnothing$.

Proof. The proof is organized in six steps. First, we follow the argument of Meinshausen which leaves us to bound the term $\left\| P_S Y - \hat{X}\beta^* \right\|_2$. In the second step, we use a decomposition proposed

by Zhu to decompose this term into three summands which are then bounded separately in step 3,4 and 5. Finally, we conclude in step 6.

Step 1: Follow Meinshausen

Consider the objective of the oracle problem (oracle-NN2SLS). It will be useful to decompose the objective into two parts.[22] Let $P_S = \hat{X}_S(\hat{X}_S'\hat{X}_S)^{-1}\hat{X}_S'$ denote the projection onto the vectorspace spanned by the columns in $\hat{X}_S$. Then we have for any $b \in \mathbb{R}_+^s$

$$\begin{aligned}\left\|Y - \hat{X}_s b\right\|_2^2 &= \left\|P_S\left(Y - \hat{X}_s b\right)\right\|_2^2 + \left\|(I - P_S)\left(Y - \hat{X}_s b\right)\right\|_2^2 \\ &= \left\|P_S Y - \hat{X}_s b\right\|_2^2 + \left\|(I - P_S)\, Y\right\|_2^2.\end{aligned}$$

As $\|(I - P_S)\, Y\|_2^2$ does not depend on b, solving (oracle-NN2SLS) is equivalent to solving

$$\min_{b \in \mathbb{R}_+^s} \left\|P_S Y - \hat{X}_S b\right\|_2^2. \tag{2.26}$$

Note that β_S^* is feasible in (2.26). In particular, we have

$$\left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2^2 \geqslant \left\|P_S Y - \hat{X}_S \hat{\beta}^{oracle}\right\|_2^2. \tag{2.27}$$

Note that by the reverse triangle inequality, we have

$$\begin{aligned}&\left\|P_S Y - \hat{X}_S \hat{\beta}_S^{\text{oracle}}\right\|_2^2 \\ \geqslant &\left(\left\|\hat{X}_S \beta_S^* - \hat{X}_S \hat{\beta}_S^{\text{oracle}}\right\|_2 - \left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2\right)^2 \\ = &\left\|\hat{X}_2 \beta_2^* - \hat{X}_S \hat{\beta}_S^{\text{oracle}}\right\|_2^2 - 2\left\|\hat{X}_S \beta_S^* - \hat{X}_S \hat{\beta}_S^{\text{oracle}}\right\|_2 \left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2 + \left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2^2.\end{aligned}$$

[22] We explain the benefits in step 3.

Combining these two inequalities, we find

$$\begin{aligned}&\left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2^2 \\ \geqslant &\left\|\hat{X}_S \beta_S^* - \hat{X}_S \hat{\beta}_S^{\text{oracle}}\right\|_2^2 - 2\left\|\hat{X}_S \beta_S^* - \hat{X}_S \hat{\beta}_S^{\text{oracle}}\right\|_2 \left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2 + \left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2^2.\end{aligned}$$

Rearranging, we find

$$2\left\|\hat{X}_S\left(\beta_S^* - \hat{\beta}_S^{\text{oracle}}\right)\right\|\left\|P_S Y - \hat{X}_S \beta_S^*\right\| \geqslant \left\|\hat{X}_S\left(\beta_S^* - \hat{\beta}_S^{\text{oracle}}\right)\right\|_2^2.$$

Dividing by $\left\|\hat{X}_S\left(\beta_S^* - \hat{\beta}_S^{\text{oracle}}\right)\right\|$ yields

$$\left\|\hat{X}_S\left(\hat{\beta}_S^{\text{oracle}} - \beta_S^*\right)\right\|_2 \leqslant 2\left\|P_S Y - \hat{X}_S \beta_S^*\right\|_2.$$

Now we can use Assumption 5, specifically (2.17) to bound the term on the left hand side from below:

$$\sqrt{\frac{\phi_\infty n}{s}}\left\|\beta_S^* - \hat{\beta}^{\text{oracle}}\right\|_1 \leqslant \left\|\hat{X}_S\left(\hat{\beta}_S^{\text{oracle}} - \beta_S^*\right)\right\|_2.$$

To conclude, we have found

$$\left\|\beta^* - \hat{\beta}^{\text{oracle}}\right\|_1 \leqslant 2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S Y - \hat{X}\beta^*\right\|_2. \tag{2.28}$$

So the remainder of the proof has the objective of bounding the term $\left\|P_S Y - \hat{X}\beta^*\right\|_2$ from above.

Step 2: Zhu's Decomposition

We use a decomposition due to Zhu[23]. For this, recall

$$\xi := (X^* - \hat{X})\beta^* + \eta\beta^* + \varepsilon.$$

Noting that $\hat{X}\beta^* = \hat{X}_S\beta^*_S$, we have

$$\begin{aligned}
\left\|P_S Y - \hat{X}_S\beta^*_S\right\|_2 &= \left\|P_S(\hat{X}_S\beta^*_S + \xi) - \hat{X}_S\beta^*_S\right\|_2 \\
&= \left\|P_S\xi\right\|_2 \\
&= \left\|P_S(\hat{X} - X^*)\beta + P_S\eta\beta^* + P_S\varepsilon\right\|_2 \\
&\leqslant \left\|P_S(\hat{X} - X^*)\beta\right\|_2 + \left\|P_S\eta\beta^*\right\|_2 + \left\|P_S\varepsilon\right\|_2 \\
&= \left\|P_S(\hat{X}_S - X^*_S)\beta^*_S\right\|_2 + \left\|P_S\eta\beta^*\right\|_2 + \left\|P_S\varepsilon\right\|_2 . \qquad (2.29)
\end{aligned}$$

The term $\left\|P_S\varepsilon\right\|_2$ also appears in the analysis of Meinshausen[24]. Compared to Meinshausen's analysis, endogeneity leaves us with two additional terms, $\left\|(\hat{X}_S - X^*_S)\beta^*\right\|_2$ and $\left\|P_S\eta\beta^*\right\|_2$: If there is no endogeneity, then $Z = X$ is a valid instrument so that the first stage errors η are equal to zero and hence $\hat{X} = X^*$. So additional terms vanish when there is no endogeneity.

Combining (2.29) with (2.28), we find

$$\left\|\beta^* - \hat{\beta}^{\text{oracle}}\right\|_1 \leqslant 2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S(\hat{X}_S - X^*_S)\beta^*_S\right\|_2 + 2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S\eta\beta^*\right\|_2 + 2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S\varepsilon\right\|_2 . \qquad (2.30)$$

[23] See the proof of Lemma A.1 in Zhu (2018).
[24] See Lemma 4 in Meinshausen (2013).

In the following three steps, we derive bounds for each of the three summands appearing on the right hand side of inequality (2.30).

Step 3: Bound $2\sqrt{\frac{\mathbf{s}}{\mathbf{n}\phi_\infty}}\left\|\mathbf{P_S\varepsilon}\right\|_\mathbf{2}$

Recall that the the term $\left\|P_S\varepsilon\right\|_2$ also appears in the exogenous case considered by Meinshausen. We find it instructive to highlight some similarities and differences between the lines of arguments of Meinhausen and our paper.

Meinshausen assumes Gaussianity of ε and that X is fixed, so the term $\left\|P_S\varepsilon\right\|_2$ follows a $\chi^2(s)$ distribution and exact tail probabilities can be derived. [25] In our setting, two complications arise.

First, we do not assume exact Gaussianity but only sub-Gaussianity. As a result, we will not be able to obtain exact tail probabilities. Instead, we will derive bounds on tail probabilities which will depend on constants from Hoeffding's and Bernstein's concentration inequality.

The second point is more substantial. As we study endogeneity, we have to allow for randomness in Z, X and hence $\hat{X}$. Endogeneity also requires that we allow for correlation between η and ε so that $\hat{X}$ and ε are not independent of one another. This makes deriving tail bounds for $\left\|P_S\varepsilon\right\|_2^2$ more challenging: Even if we were to assume Gaussianity of ε, we would not be able to infer that $\hat{x}_j'\varepsilon$ follows a Gaussian distribution.

P_S is a projection onto the column space generated by $\hat{X}_S$ so that it has the eigenvalue 1 with multiplicity s and the eigenvalue 0 with multiplicity $n-s$ because we assume that $\hat{X}_S$ has full rank per Assumption 5, specifically (2.17). Because P_S is also symmetric, it admits an orthogo-

[25] Note that we could have done Step 1 without the projection P_S in which case we would now have $\left\|\varepsilon\right\|_2^2$. That would be a $\chi^2(n)$ in the Gaussian case so that we would not have convergence of $\sqrt{\frac{s}{n}}\left\|P_SY - \hat{X}\beta^*\right\|_2$ to zero. As a result, we would also not get convergence of $\hat{\beta}^{\text{oracle}}$ to β^*.

nal diagonalization, i.e. there exists an orthogonal matrix V such that $P_S = VDV^t$ where D is a diagonal matrix whose first s entries on the diagonal are 1 and whose last $n - s$ entries are 0. Column j of matrix V is the eigenvector associated with the j-th element on the diagonal of D. In particular, the first s columns of V are the normalized columns of $\hat{X}$ because P_S is the projection on the column space of $\hat{X}$, so that the eigenvectors corresponding to the eigenvalues 1 are just the normalized columns of $\hat{X}_S$ because the column vectors of $\hat{X}_S$ trivially span the column space of $\hat{X}_S$. Hence

$$2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S\varepsilon\right\|_2^2 = 2\sqrt{\frac{s}{n\phi_\infty}}\left\|(VDV^t)\varepsilon\right\|_2^2 = 2\sqrt{\frac{s}{n\phi_\infty}}\left\|DV^t\varepsilon\right\|_2^2. \qquad (V \text{ orthognonal})$$

We further have

$$2\sqrt{\frac{s}{n\phi_\infty}}\left\|DV^t\varepsilon\right\|_2^2 = 2\sqrt{\frac{s}{n\phi_\infty}}\sum_{j=1}^{s}\left(\frac{\hat{x}_j'\varepsilon}{\|\hat{x}_j\|}\right)^2 \leqslant 2s\sqrt{\frac{s}{n\phi_\infty}}\max_{j=1,\ldots,s}\left(\frac{\hat{x}_j'\varepsilon}{\|\hat{x}_j\|}\right)^2. \qquad (2.31)$$

If $\hat{x}_j$ were independent of ε, we could apply Hoeffding's inequality conditionally on $\hat{x}_j$ and then note that the bound is independent of $\hat{x}_j$. However, $\hat{x}_j$ depends on η which we expect to be correlated with ε. So we have to show that $\frac{\hat{x}_j'\varepsilon}{\|\hat{x}_j\|}$ can be bounded with high probability. We have

$$2s\sqrt{\frac{s}{n\phi_\infty}}\varepsilon'\frac{\hat{x}_j}{\|\hat{x}_j\|_2} = 2s\sqrt{\frac{s}{\phi_\infty}}\frac{1}{\frac{1}{\sqrt{n}}\|\hat{x}_j\|_2}\left(\frac{1}{n}\varepsilon' x_j^* + \frac{1}{n}\varepsilon'(\hat{x}_j - x_j^*)\right) \qquad (2.32)$$

We now bound the right hand side from above by bounding each of the three terms separately:

1. the term $\frac{1}{\frac{1}{\sqrt{n}}\|\hat{x}_j\|_2}$,
2. the term $\frac{1}{n}\varepsilon' x_j^*$,
3. the term $\frac{1}{n}\varepsilon'(\hat{x}_j - x_j^*)$.

Bounding the term $\frac{1}{\frac{1}{\sqrt{n}}\|\hat{x}_j\|_2}$ Our strategy will be to bound $\frac{1}{\sqrt{n}}\|\hat{x}_j\|_2$ from below. For this, we use Lemma 45 to infer that with probability at least $p_{L45}(S)$ we have that for all $j \in S$

$$\frac{1}{\sqrt{n}}\|\hat{x}_j\|_2 \geqslant r_{L45}(j),$$

where $p_{L45}(\cdot)$ is defined in (2.43) and $r_{L45}(\cdot)$ is defined in (2.44). Hence with probability at least $p_{L45}(S)$ we have that for all $j \in S$

$$\frac{1}{\frac{1}{\sqrt{n}}\|\hat{x}_j\|_2} \leqslant \frac{1}{r_{L45}(j)}.$$

Bounding the term $\frac{1}{n}\varepsilon' x_j^*$ We have

$$\begin{aligned}
\frac{1}{n}\varepsilon' x_j^* =& \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_{i,j}^* \\
=& \frac{1}{n}\sum_{i=1}^n \varepsilon_i \left(\sum_{l=1}^d Z_{i,l}\pi_{l,j}^*\right) \\
=& \frac{1}{n}\sum_{l=1}^d \pi_{l,j}^* \left(\sum_{i=1}^n \varepsilon_i Z_{i,\cdot}\right) \\
\leqslant& \|\pi_j^*\|_1 \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i Z_{i,\cdot}\right\|_\infty .
\end{aligned}$$

Recall that by Assumption 3, $\varepsilon_i Z_{i,l}$ is independent over i and has mean zero. Because both ε_i and $Z_{i,l}$ are sub-Gaussian by Assumption 4, their product is sub-exponential with norm at most $\rho_\varepsilon \rho_Z$. Hence we can apply Bernstein's inequality[26] to infer that a fixed $l \in \{1, \ldots, d\}$ we have

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_i Z_{i,l}\right| > t\right] \leqslant 2\exp\left(-c_B \min\left\{\frac{t}{\rho_z\rho_\varepsilon}, \frac{t^2}{\rho_z^2\rho_\varepsilon^2}\right\} n\right).$$

[26]Corollary 2.8.3 in Vershynin (2020).

In particular, if we choose $t = \frac{1}{n^\gamma}\rho_z\rho_\varepsilon$, then with probability at least $1 - 2\exp\left(-c_B n^{1-2\gamma}\right)$ we have

$$\frac{1}{n}\left|\sum_{i=1}^{n}\varepsilon_i Z_{i,l}\right| \leqslant \frac{1}{n^\gamma}\rho_z\rho_\varepsilon.$$

Using a union bound, we see that with probability at least $1 - 2d\exp\left(-c_B n^{1-2\gamma}\right)$ we have

$$\max_{l\in\{1,\ldots,d\}}\frac{1}{n}\left|\sum_{i=1}^{n}\varepsilon_i Z_{i,l}\right| \leqslant \frac{1}{n^\gamma}\rho_z\rho_\varepsilon.$$

To summarize, we have shown that with probability at least $1 - 2d\exp\left(-c_B n^{1-2\gamma}\right)$ it holds that

$$\frac{1}{n}\varepsilon' x_j^* \leqslant \left\|\pi_j^*\right\|_1 \frac{1}{n^\gamma}\rho_z\rho_\varepsilon.$$

Bounding the term $\frac{1}{n}\varepsilon'(\hat{x}_j - x_j^*)$ We have

$$\frac{1}{n}\varepsilon'(\hat{x}_j - x_j^*) = \frac{1}{n}\varepsilon' Z\left(\hat{\pi}_j - \pi_j^*\right) \leqslant \frac{1}{\sqrt{n}}\left\|\varepsilon\right\|_2 \frac{1}{\sqrt{n}}\left\|Z\left(\hat{\pi}_j - \pi_j^*\right)\right\|_2.$$

By assumption 4, ε_i is sub-Gaussian with parameter ρ_ε. By (Vershynin, 2020, Lemma 2.7.6), ε_i^2 is sub-exponential with constant at most ρ_ε^2. So when we subtract the mean of ε_i^2, it is sub-exponential with constant at most $c_{center,SE}\varepsilon_i^2$. Because we have assumed independence over i in assumption 3, we can apply Bernstein's inequality (Vershynin, 2020, Corollary 2.8.3) to see

$$\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\left(\left\|\varepsilon\right\|_2^2 - \mathbb{E}[\left\|\varepsilon\right\|_2^2]\right)\right| \geqslant t\right] \leqslant 2\exp\left(-c_B\min\left\{\frac{t^2}{c_{center,SE}^2\rho_\varepsilon^4}, \frac{t}{c_{center,SE}\rho_\varepsilon^2}\right\}n\right).$$

For the second term, we can use Assumption 6 to infer that with probability at least $p_{A4,j,n}$ we have that $\frac{1}{\sqrt{n}}\left\|Z\left(\hat{\pi}_j - \pi_j^*\right)\right\|_2 \leqslant r_{A4,j,n}$. Combining the bounds for the two terms with a union bound,

we see that with probability at least $p_{A4,j,n} - 2\exp(-c_B n)$ it holds that

$$\frac{1}{n}\varepsilon'(\hat{x}_j - x_j^*) \leqslant \sqrt{2c_{center,SE}}\rho_\varepsilon r_{A4,j,n}.$$

To summarize, we combine (2.31), (2.32) and our probabilistic bounds for the three terms appearing in (2.32) with a union bound to infer that

$$2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S\varepsilon\right\|_2 \leqslant 2s\sqrt{\frac{s}{\phi_\infty}}\frac{1}{\min\limits_{j\in S} r_{L45}(j)}\left(\frac{\rho_Z\rho_\varepsilon}{n^\gamma}\max_{j\in S}\left\|\pi_j^*\right\|_1 + \sqrt{2c_{center,SE}}\rho_\varepsilon r_{A4,S,n}\right)$$

with probability at least

$$p_{S3,\gamma} := p_{A4,S,n} + p_{L45}(S) - 1 - 2\exp(-c_B n) - 2d\exp(-c_B n^{1-2\gamma}). \tag{2.33}$$

Step 4: Bound $2\sqrt{\frac{\mathbf{s}}{\mathbf{n}\phi_\infty}}\left\|\mathbf{P_S}\eta\beta^*\right\|_\mathbf{2}^\mathbf{2}$

Note that $\eta\beta^*$ is a vector with n i.i.d. entries. By Assumption 4, we know that $\sum\limits_{j\in S}\eta_{1,j}\beta_j^*$ is sub-Gaussian with sub-Gaussian norm at most $\rho_\eta\left\|\beta^*\right\|_2$. Hence the argument to bound $\left\|P_S\eta\beta^*\right\|_2^2$ proceeds exactly as in step 3, we just have to replace ρ_ε with $\rho_\eta\left\|\beta\right\|_2$.

Step 5: Bound $2\sqrt{\frac{\mathbf{s}}{\mathbf{n}\phi_\infty}}\left\|\mathbf{P_S}(\hat{\mathbf{X}}_\mathbf{S} - \mathbf{X}_\mathbf{S}^*)\boldsymbol{\beta}^*\right\|_\mathbf{2}^\mathbf{2}$

Note that

$$\begin{aligned}
2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S(\hat{X}_S - X_S^*)\beta_S^*\right\|_2 &\leqslant 2\sqrt{\frac{s}{n\phi_\infty}}\left\|(\hat{X}_S - X_S^*)\beta_S^*\right\|_2 && (P_S \text{ is a projection})\\
&= 2\sqrt{\frac{s}{n\phi_\infty}}\left\|Z(\hat{\pi}_S - \pi_S^*)\beta_S^*\right\|_2\\
&= 2\sqrt{\frac{s}{n\phi_\infty}}\left\|\sum_{j\in S} Z(\hat{\pi}_j - \pi_j^*)\beta_j^*\right\|_2\\
&\leqslant 2\sqrt{\frac{s}{n\phi_\infty}}\sum_{j\in S}\left\|Z(\hat{\pi}_j - \pi_j^*)\beta_j^*\right\|_2
\end{aligned}$$

$$\leqslant 2s\sqrt{\frac{s}{n\phi_\infty}}\max_{j\in S}\left\|Z(\hat{\pi}_j-\pi_j^*)\beta_j^*\right\|_2$$
$$\leqslant 2s\sqrt{\frac{s}{\phi_\infty}}\left\|\beta^*\right\|_\infty\frac{1}{\sqrt{n}}\max_{j\in S}\left\|Z(\hat{\pi}_j-\pi_j^*)\right\|_2.$$

We are now in the position to apply Assumption 6 to bound $\max_{j\in S}\left\|Z(\hat{\pi}_j-\pi_j^*)\right\|_\infty$. Hence we conclude that with probability at least $p_{A4,S,n}$

$$2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_S(\hat{X}_S-X_S^*)\beta^*\right\|_2\leqslant s2\sqrt{\frac{s}{\phi_\infty}}\left\|\beta^*\right\|_\infty r_{A4,S,n}.$$

Step 6: Conclude

Let's recap the results from step 1 and 2:

$$\left\|\beta^*-\hat{\beta}^{\text{oracle}}\right\|_1\leqslant 2\sqrt{\frac{s}{n\phi_\infty}}\left\|P_SY-\hat{X}\beta^*\right\|_2 \tag{(2.28)}$$
$$\leqslant 2\sqrt{\frac{s}{n\phi_\infty}}\left(\left\|P_S(\hat{X}_S-X_S^*)\beta^*\right\|_2+\left\|P_S\eta\beta^*\right\|_2+\left\|P_S\varepsilon\right\|_2\right). \tag{(2.30)}$$

To conclude, we use a union bound to combine the bounds from steps 3,4 and 5. □

Lemma 43. Assume that Assumptions 3, 4 and 6 are satisfied. Fix $\alpha\in\left(0,\frac{1}{2}\right)$. Define $\xi\in\mathbb{R}^n$ as

$$\xi:=(\hat{X}-X^*)\beta+\eta\beta^*+\varepsilon. \tag{2.34}$$

Then with probability at least

$$p_{L4}:=p_{S2}+p_{S3}+p_{S4}(\alpha)-2$$

it holds that

$$\left\|\frac{1}{n}\hat{X}^t\xi\right\|_\infty\leqslant r_{S2}+r_{S3}+r_{S4}(\alpha):=r_{L4} \tag{2.35}$$

where p_{S2} is defined in (2.37), r_{S2} is defined in (2.38), p_{S3} is defined in (2.39), r_{S3} is defined in (2.40), $p_{S4}(\alpha)$ is defined in (2.41), $r_{S4}(\alpha)$ is defined in (2.42).

Proof. This proof is similar to proof of Lemma A.4 in Zhu (2018). First, the triangle inequality gives

$$\begin{aligned}
&\left\|\frac{1}{n}\hat{X}^t\xi\right\|_\infty \\
=&\left\|\frac{1}{n}\hat{X}^t\left((\hat{X}-X^*)\beta+\eta\beta^*+\varepsilon\right)\right\|_\infty \\
\leqslant&\left\|\frac{1}{n}\hat{X}^t(\hat{X}-X^*)\beta\right\|_\infty+\frac{1}{n}\left\|\hat{X}^t\left(\eta\beta^*+\varepsilon\right)\right\|_\infty \\
\leqslant&\left\|\frac{1}{n}\hat{X}^t(\hat{X}-X^*)\beta\right\|_\infty+\frac{1}{n}\left\|\left(\hat{X}-X^*\right)^t\left(\eta\beta^*+\varepsilon\right)\right\|_\infty+\frac{1}{n}\left\|(X^*)^t\left(\eta\beta^*+\varepsilon\right)\right\|_\infty \\
=&\left\|\frac{1}{n}\hat{X}^t(\hat{X}-X^*)\beta\right\|_\infty+\frac{1}{n}\left\|\left(\hat{X}-X^*\right)^t\left(\eta_S\beta_S^*+\varepsilon\right)\right\|_\infty+\frac{1}{n}\left\|(X^*)^t\left(\eta_S\beta_S^*+\varepsilon\right)\right\|_\infty. \qquad (2.36)
\end{aligned}$$

In the following three steps, we will bound each of these terms separately.

Step 2: Bound $\left\|\frac{1}{n}\hat{X}(\hat{X}-X^*)^t\beta\right\|_\infty$

Recall that

$$\left\|\frac{1}{n}\hat{X}'(\hat{X}-X^*)\beta^*\right\|_\infty=\max_{k\in\{1,\ldots,p\}}\left|\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^*\right|.$$

Now consider an arbitrary $k\in\{1,\ldots,p\}$. Then

$$\begin{aligned}
&\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^* \\
=&\frac{1}{n}\hat{X}_k'Z(\hat{\pi}-\pi^*)\beta^* \\
=&\frac{1}{n}\hat{X}_k'Z(\hat{\pi}_S-\pi_S^*)\beta_S^* \\
\leqslant&\frac{1}{n}\left\|\hat{X}_k\right\|_2\left\|Z(\hat{\pi}_S-\pi_S^*)\beta_S^*\right\|_2 \qquad \text{(Hölder)}
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{n}\left\|\hat{X}_k\right\|_2\left\|\sum_{j\in S} Z(\hat{\pi}_j-\pi_j^*)\beta_j^*\right\|_2\\
\leqslant&\frac{s}{n}\left\|\hat{X}_k\right\|_2\|\beta^*\|_\infty \max_{j\in S}\left\|Z(\hat{\pi}_j-\pi_j^*)\right\|_2\\
=&s\frac{1}{\sqrt{n}}\left\|\hat{X}_k\right\|_2\|\beta^*\|_\infty \frac{1}{\sqrt{n}}\max_{j\in S}\left\|Z(\hat{\pi}_j-\pi_j^*)\right\|_2\\
\leqslant&sr_{L45}(k)\|\beta^*\|_\infty \frac{1}{\sqrt{n}}\max_{j\in S}\left\|Z(\hat{\pi}_j-\pi_j^*)\right\|_2 && \text{(w.p. } \geqslant p_{L45}(\{k\}) \text{ by Lemma 45)}\\
\leqslant&sr_{L45}(k)\|\beta^*\|_\infty r_{A4,S,n}, && \text{(w.p. } \geqslant p_{A4,S,n} \text{ by A6)}
\end{aligned}
$$

where p_{L45} is defined in 2.43 and r_{L45} is defined in 2.44. With a union bound, we conclude that with probability at least

$$p_{L45}(\{k\})+p_{A4,S,n}-1$$

it holds that

$$\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^* \leqslant sr_{L45}(k)\|\beta^*\|_\infty r_{A4,S,n} =: b_{n,k}.$$

Now let's use a union bound to control the maximum over $k\in\{1,\dots,p\}$:

$$
\begin{aligned}
&\mathbb{P}\left[\max_{k\in\{1,\dots,p\}}\left|\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^*\right| \geqslant \max_{k\in\{1,\dots,p\}} b_{n,k}\right]\\
=&\mathbb{P}\left[\bigcup_{k\in\{1,\dots,p\}}\left\{\left|\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^*\right| \geqslant \max_{k\in\{1,\dots,p\}} b_{n,k}\right\}\right]\\
\leqslant&\sum_{k\in\{1,\dots,p\}}\mathbb{P}\left[\left|\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^*\right| \geqslant \max_{k\in\{1,\dots,p\}} b_{n,k}\right]\\
\leqslant&\sum_{k\in\{1,\dots,p\}}\mathbb{P}\left[\left|\frac{1}{n}\hat{X}_k'(\hat{X}-X^*)\beta^*\right| \geqslant b_{n,k}\right]\\
\leqslant&1-p_{A4,S,n}-p_{L45}(\{1,\dots,p\}).
\end{aligned}
$$

Concluding with a union bound, with probability at least

$$p_{S2} := p_{A4,S,n} + p_{L45}(\{1,\ldots,p\}) - 1 \tag{2.37}$$

it holds that

$$\left\| \frac{1}{n} \hat{X}'(\hat{X} - X^*)\beta^* \right\|_\infty \leqslant s \left\|\beta^*\right\|_\infty r_{A4,S,n} \left(\max_{k \in \{1,\ldots,p\}} r_{L45}(k) \right) =: r_{S2}. \tag{2.38}$$

Step 3: Bound $\frac{1}{n} \left\| \left(\hat{X} - X^*\right)^t (\eta_S \beta_S^* + \varepsilon) \right\|_\infty$

We have

$$\begin{aligned}
&\frac{1}{n} \left\| \left(\hat{X} - X^*\right)' (\eta_S \beta_S^* + \varepsilon) \right\|_\infty \\
=&\frac{1}{n} \left\| Z \left(\hat{\pi} - \pi^*\right) (\eta_S \beta_S^* + \varepsilon) \right\|_\infty \\
=& \max_{j \in \{1,\ldots,p\}} \frac{1}{n} \left| \left(\hat{\pi}_j - \pi_j^*\right)' Z' (\eta_S \beta_S^* + \varepsilon) \right| \\
\leqslant& \max_{j \in \{1,\ldots,p\}} \frac{1}{n} \left\| \left(\hat{\pi}_j - \pi_j^*\right)' Z' \right\|_2 \left\| \eta_S \beta_S^* + \varepsilon \right\|_2 && \text{(Hölder)} \\
=&\frac{1}{\sqrt{n}} \left\| \eta_S \beta_S^* + \varepsilon \right\|_2 \max_{j \in \{1,\ldots,p\}} \frac{1}{\sqrt{n}} \left\| Z \left(\hat{\pi}_j - \pi_j^*\right) \right\|_2 \\
\leqslant&\frac{1}{\sqrt{n}} \left(\left\| \eta_S \beta_S^* \right\|_2 + \left\| \varepsilon \right\|_2 \right) \max_{j \in \{1,\ldots,p\}} \frac{1}{\sqrt{n}} \left\| Z \left(\hat{\pi}_j - \pi_j^*\right) \right\|_2 .
\end{aligned}$$

We can control $\left\|\eta_S \beta_S^*\right\|_2$ and $\left\|\varepsilon\right\|_2$ via Bernstein's inequality: By assumption 4, $\eta_{i,S}\beta_S^*$ is sub-Gaussian with parameter $\rho_\eta \left\|\beta_S^*\right\|_2$. By (Vershynin, 2020, Lemma 2.7.6), $\left(\eta_{i,S}\beta_S^*\right)^2$ is sub-exponential with constant at most $\rho_\eta^2 \left\|\beta_S^*\right\|_2^2$. So when we subtract the mean of $\left(\eta_{i,S}\beta_S^*\right)^2$, it is sub-exponential with constant at most $c_{center,SE} \left(\eta_{i,S}\beta_S^*\right)^2$. Because we have assumed independence over i in assumption 3, we can apply Bernstein's inequality (Vershynin, 2020, Corollary 2.8.3). Using the

same arguments for $\|\varepsilon\|_2$, we find

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\left|\|\eta_S\beta_S^*\|_2^2 - \mathbb{E}\left[\|\eta_S\beta_S^*\|_2^2\right]\right| \geqslant t\right] \\
\leqslant& 2\exp\left(-c_B\min\left\{\frac{t^2}{c_{center,SE}^2\rho_\eta^4\|\beta^*\|_2^4}, \frac{t}{c_{center,SE}\rho_\eta^2\|\beta^*\|_2^2}\right\}n\right), \\
&\mathbb{P}\left[\frac{1}{n}\left|\|\varepsilon\|_2^2 - \mathbb{E}\left[\|\varepsilon\|_2^2\right]\right| \geqslant t\right] \\
\leqslant& 2\exp\left(-c_B\min\left\{\frac{t^2}{c_{center,SE}^2\rho_\varepsilon^4}, \frac{t}{c_{center,SE}\rho_\varepsilon^2}\right\}n\right).
\end{aligned}$$

With Assumption 6 and a union bound, we see that with probability at least

$$p_{S3} := p_{A4,\{1,\ldots,p\},n} - 4\exp(-c_B n) \tag{2.39}$$

it holds that

$$\frac{1}{n}\left\|\left(\hat{X} - X^*\right)'(\eta_S\beta_S^* + \varepsilon)\right\|_\infty \leqslant \sqrt{2c_{center,SE}}\left(\rho_\eta\|\beta^*\|_2 + \rho_\varepsilon\right) r_{A4,\{1,\ldots,p\},n} =: r_{S3}. \tag{2.40}$$

Step 4: Bound $\frac{1}{n}\|X^*(\eta_S\beta_S^* + \varepsilon)\|_\infty$

We have

$$\begin{aligned}
&\frac{1}{n}\left\|(X^*)^t(\eta_S\beta_S^* + \varepsilon)\right\|_\infty \\
=& \max_{j\in\{1,\ldots,p\}}\frac{1}{n}\left|\left(X_j^*\right)^t(\eta_S\beta_S^* + \varepsilon)\right| \\
=& \max_{j\in\{1,\ldots,p\}}\frac{1}{n}\left|\left(Z\pi_j^*\right)^t(\eta_S\beta_S^* + \varepsilon)\right| \\
=& \max_{j\in\{1,\ldots,p\}}\frac{1}{n}\left|\left(\pi_j^*\right)^t Z^t(\eta_S\beta_S^* + \varepsilon)\right| \\
=& \max_{j\in\{1,\ldots,p\}}\left|\frac{1}{n}\sum_{i=1}^n\left(\sum_{l=1}^d \pi_{l,j}^* Z_{i,l}\right)\left(\sum_{k\in S}\eta_{i,k}\beta_k^* + \varepsilon_i\right)\right|
\end{aligned}$$

$$\leqslant \max_{j\in\{1,\ldots,p\}} \left(\left| \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{l=1}^{d} \pi^*_{l,j} Z_{i,l} \right) \left(\sum_{k\in S} \eta_{i,k}\beta^*_k \right) \right| + \left| \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{l=1}^{d} \pi^*_{l,j} Z_{i,l} \right) \varepsilon_i \right| \right)$$

By Assumption 4, $\sum_{l=1}^{d} \pi^*_{l,j} Z_{i,l}$ is sub-Gaussian with sub-Gaussian norm at most $\rho_z \left\|\pi^*_j\right\|_2$. Similarly, $\sum_{k\in S} \eta_{i,k}\beta^*_k$ is sub-Gaussian with sub-Gaussian norm at most $\rho_\eta \left\|\beta^*\right\|_2$. Now we can use Lemma 2.7.7 in Vershynin (2020) infer that

$$W^1_{i,j} := \left(\sum_{l=1}^{d} \pi^*_{l,j} Z_{i,l} \right) \left(\sum_{k\in S} \eta_{i,k}\beta^*_k \right)$$
$$W^2_{i,j} := \left(\sum_{l=1}^{d} \pi^*_{l,j} Z_{i,l} \right) \varepsilon_i$$

are sub-exponential with sub-exponential with sub-exponential norm at most

$$\rho_{W^1,j} = \rho_z \left\|\pi^*_j\right\|_2 \rho_\eta \left\|\beta^*\right\|_2$$
$$\rho_{W^2,j} = \rho_z \left\|\pi^*_j\right\|_2 \rho_\varepsilon,$$

respectively. By Assumption 3, we also know that W^1_i and W^2_i have expectation zero. Also by Assumption 3, $(W^1_{i,j})$ is independent over i, as is $(W^2_{i,j})$. Hence we can apply Bernstein's inequality to infer for $k \in \{1, 2\}$

$$\mathbb{P}\left[\left| \frac{1}{n}\sum_{i=1}^{n} W^k_{i,j} \right| \geqslant t \right] \leqslant 2\exp\left(-c_B \min\left\{ \frac{t^2}{\rho^2_{W^k,j}}, \frac{t}{\rho_{W^k,j}} \right\} n \right).$$

With $t = \max_{j\in\{1,\ldots,p\}} \frac{\rho_{W^k,j}}{n^\alpha}$, a union bound over j yields

$$\mathbb{P}\left[\max_{j\in\{1,\ldots,p\}} \left| \frac{1}{n}\sum_{i=1}^{n} W^k_{i,j} \right| \geqslant \frac{\max_{j\in\{1,\ldots,p\}} \rho_{W,j}}{n^\alpha} \right] \leqslant 2p\exp\left(-c_B n^{1-2\alpha} \right).$$

Taking a union bound over $k \in \{1, 2\}$ yields that with probability at least

$$p_{S4}(\alpha) := 1 - 4p \exp\left(-c_B n^{1-2\alpha}\right) \tag{2.41}$$

it holds that

$$\frac{1}{n} \left\| (X^*)^t \left(\eta_S \beta_S^* + \varepsilon\right) \right\|_\infty \leqslant \frac{1}{n^\alpha} \max_{j \in \{1,\ldots,p\}} \rho_z \left\| \pi_j^* \right\|_2 \left(\rho_\eta \left\| \beta^* \right\|_2 + \rho_\varepsilon\right) =: r_{S4}(\alpha). \tag{2.42}$$

Step 5: Conclude

Using a union bound, we combine (2.38), (2.40) and (2.42) to bound the three terms in (2.36). □

Lemma 44. Suppose Assumption 3, 4 and 6 are satisfied. Then with probability at least $p_{L45}(S)$ it holds that

$$\sigma_{\max}\left(\frac{\hat{X}_S^t \hat{X}_S}{n}\right) \leqslant s \left(\max_{j \in S} r_{L45}(j)\right)^2.$$

where $p_{L45}(S)$ is defined in (2.43) and $r_{L45}(S)$ is defined in (2.44).

Proof. Recall that

$$\begin{aligned}
\sigma_{\max}\left(\frac{\hat{X}_S^t \hat{X}_S}{n}\right) &\leqslant \frac{1}{n} \max_{j \in S} \sum_{k \in S} \left| \hat{X}_j' \hat{X}_k \right| && \text{(Gershgorin)} \\
&\leqslant \frac{1}{n} \max_{j \in S} \sum_{k \in S} \left\| \hat{X}_j \right\|_2 \left\| \hat{X}_k \right\|_2 && \text{(Hölder)} \\
&= \frac{1}{n} \max_{j \in S} \left\| \hat{X}_j \right\|_2 \sum_{k \in S} \left\| \hat{X}_k \right\|_2 \\
&\leqslant s \left(\max_{j \in S} \frac{1}{\sqrt{n}} \left\| \hat{X}_j \right\|_2 \right)^2 \\
&\leqslant s \left(\max_{j \in S} r_{L45}(j) \right)^2 && \text{(Lemma 45)}
\end{aligned}$$

with probability at least $p_{L45}(S)$. □

Lemma 45. Under Assumption 3, 4 and 6 for any set $\mathfrak{J} \subset \{1, \ldots, p\}$, we have with probability a least

$$p_{L45}(\mathfrak{J}) := p_{A4,\mathfrak{J},n} - 2\sum_{j\in\mathfrak{J}} \exp\left(-c_B \min\left\{\kappa_j, \kappa_j^2\right\} n\right) \tag{2.43}$$

that for all $j \in \mathfrak{J}$

$$\frac{1}{\sqrt{n}}\left\|\hat{X}_j\right\|_2 \in \left[\underbrace{\sqrt{\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]} - r_{A4,\mathfrak{J},n}}_{:=r_{L45,low}}, \underbrace{\sqrt{\frac{3}{2n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]} + r_{A4,\mathfrak{J},n}}_{:=r_{L45,high}}\right] \tag{2.44}$$

where κ_j is defined in (2.47).

Proof. Note that

$$\frac{1}{\sqrt{n}}\left\|\hat{X}_j\right\|_2 = \frac{1}{\sqrt{n}}\left\|Z\hat{\pi}_j\right\|_2 = \frac{1}{\sqrt{n}}\left\|Z\pi_j^* - Z(\pi_j^* - \hat{\pi}_j)\right\|_2.$$

The (reverse) triangle inequality yields

$$\frac{1}{\sqrt{n}}\left\|Z\pi_j^*\right\|_2 - \frac{1}{\sqrt{n}}\left\|Z\left(\hat{\pi}_j - \pi_j^*\right)\right\|_2 \leqslant \frac{1}{\sqrt{n}}\left\|\hat{X}_j\right\|_2 \leqslant \frac{1}{\sqrt{n}}\left\|Z\pi_j^*\right\|_2 + \frac{1}{\sqrt{n}}\left\|Z\left(\hat{\pi}_j - \pi_j^*\right)\right\|_2. \tag{2.45}$$

Note that $Z_{i,\cdot}\pi_j^*$ is sub-Gaussian with constant at most $\rho_z\left\|\pi_j^*\right\|_2$ by Assumption 4. We have

$$\frac{1}{\sqrt{n}}\left\|Z\pi_j^*\right\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_{i,\cdot}\pi_j^*\right)^2}.$$

By (Vershynin, 2020, Lemma 2.7.6), $(Z_{i,\cdot}\pi_j^*)^2$ is sub-exponential with constant at most $\rho_z^2\left\|\pi_j^*\right\|_2^2$. So when we subtract the mean of $(Z_{i,\cdot}\pi_j^*)^2$, it is sub-exponential with constant at most

$c_{center,SE}\rho_z^2 \left\|\pi_j^*\right\|_2^2$. Because we have assumed independence over i in Assumption 3, we can apply Bernstein's inequality (Vershynin, 2020, Theorem 2.8.2) to infer that

$$\begin{aligned}&\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_{i,\cdot}\pi_j^*\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]\right| \geqslant t\right] \\ \leqslant & 2\exp\left(-c_B \min\left\{\frac{t^2}{c_{center,SE}^2\rho_Z^4\left\|\pi_j^*\right\|_2^4}, \frac{t}{c_{center,SE}\rho_Z^2\left\|\pi_j^*\right\|_2^2}\right\} n\right).\end{aligned}$$

In particular, with the choice $t = \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]$ we obtain that with probability at least $1 - 2\exp\left(-c_B \min\left\{\kappa_j, \kappa_j^2\right\} n\right)$ we have

$$\frac{1}{\sqrt{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]} \leqslant \frac{1}{\sqrt{n}}\left\|Z\pi_j^*\right\|_2 =\leqslant \sqrt{\frac{3}{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]} \tag{2.46}$$

where

$$\kappa_j := \frac{\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\left[(Z_{i,\cdot}\pi_j^*)^2\right]}{c_{center,SE}\rho_Z^2\left\|\pi_j^2\right\|_2^2}. \tag{2.47}$$

We now combine this probabilistic bound with Assumption 6 using a union bound to infer that (2.44) for all $j \in \mathfrak{J}$ with probability at least

$$1 - (1 - p_{A4,\mathfrak{J},n}) - 2\sum_{j\in\mathfrak{J}}\exp\left(-c_B \min\left\{\kappa_j, \kappa_j^2\right\} n\right) = p_{A4,\mathfrak{J},n} - \sum_{j\in\mathfrak{J}} 2\exp\left(-c_B \min\left\{\kappa_j, \kappa_j^2\right\} n\right).$$

This completes the proof. □

2.8.2 On Assumption 6

Assumption 6 is formulated for any subset $\mathfrak{J} \subset \{1, \ldots, p\}$. In contrast, results for many methods are available only for $\mathfrak{J}$ with exactly one element. The following remark show how to use a union

bound to go from such results to the one required in Assumption 6.

Remark 46. Suppose that for each $j \in \{1, \ldots, p\}$ we have that with probability at least $p_{A4,\{j\},n} =: p_{A4,j,n}$ it holds that

$$\frac{1}{\sqrt{n}} \left\| Z(\hat{\pi}_j - \pi_j^*) \right\|_2 \leqslant r_{A4,\{j\},n} =: r_{A4,j,n}.$$

Then for any $\mathfrak{J} \subset \{1, \ldots, p\}$, inequality (2.18) with $r_{A4,\mathfrak{J},n} := \max_{j \in \mathfrak{J}} r_{A4,j,n}$ holds that with probability at least

$$p_{A4,\mathfrak{J},n} := 1 - \sum_{j \in \mathfrak{J}} (1 - p_{A4,j,n}) . \tag{2.48}$$

To get some intuition, consider equation (2.48) where $p_{A4,j,n} = p_{A4,1,n}$ for all $j \in \{1, \ldots, n\}$, i.e.

$$p_{A4,\mathfrak{J},n} = 1 - |\mathfrak{J}| \, (1 - p_{A4,1,n}) .$$

For asymptotic results, it will be important that $p_{A4,\mathfrak{J},n}$ converges to 1 for any $\mathfrak{J}$, in particular for the "worst case" $\mathfrak{J} = \{1, \ldots, p\}$. For this, $(1 - p_{A4,1,n})$ will have to decrease faster than p is rising, for example exponentially in n.

2.8.3 Prediction Error of the LASSO

In this section, we revisit the analysis of the prediction error of the LASSO.

Our goal is not to generate new insights. In fact, we follow the discussion of Hastie, Tibshirani and Friedman in section 11 of Hastie et al. (2015) which, to the best of our knowledge, is based on the analysis by Bickel, Ritov and Tsybakov in Bickel et al. (2009). The motivation for including

this section is to allow the reader to follow the analysis in one coherent framework. While the literature often considers fixed design matrices and Gaussian errors, we present the results with random matrices and sub-Gaussian errors.

We start with the LASSO objective for a generic first stage $j \in \{1, \ldots, p\}$. We have

$$f(\pi_j) = \frac{1}{2n} \left\| X_j - Z\pi_j \right\|_2^2 + \lambda \left\| \pi_j \right\|_1 \tag{2.49}$$

Denote the minimizer of this function, i.e. the LASSO estimator by $\hat{\pi}_j^{\text{LASSO}}$.

Lemma 47. We have

$$\left\| \frac{1}{n} \eta_j^t Z \right\|_\infty \left\| \hat{\pi}_j^{\text{LASSO}} - \pi_j^* \right\|_1 + \lambda \left(\left\| \pi_j^* \right\|_1 - \left\| \hat{\pi}_j^{\text{LASSO}} \right\|_1 \right) \geqslant \frac{1}{2n} \left\| Z \left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^* \right) \right\|_2^2. \tag{2.50}$$

Proof. As $\hat{\pi}_j^{\text{LASSO}}$ is the minimizer of (2.49), we have

$$f\left(\pi_j^*\right) \geqslant f(\hat{\pi}_j^{\text{LASSO}}).$$

i.e.

$$\begin{aligned}
&\frac{1}{2n} \left\| X_j - Z\pi_j^* \right\|_2^2 + \lambda \left\| \pi_j^* \right\|_1 \\
\geqslant &\frac{1}{2n} \left\| X_j - Z\hat{\pi}_j^{\text{LASSO}} \right\|_2^2 + \lambda \left\| \hat{\pi}_j^{\text{LASSO}} \right\|_1 \\
= &\frac{1}{2n} \left\| X_j - Z\pi_j^* - Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^* \right) \right\|_2^2 + \lambda \left\| \hat{\pi}_j^{\text{LASSO}} \right\|_1 \\
= &\frac{1}{2n} \left(\left\| X_j - Z\pi_j^* \right\|_2^2 - 2\left(X_j - Z\pi_j^* \right)^t Z \left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^* \right) + \left\| Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^* \right) \right\|_2^2 \right) + \lambda \left\| \hat{\pi}_j^{\text{LASSO}} \right\|_1
\end{aligned}$$

Subtracting $\frac{1}{2n} \left\| X_j - Z\pi_j^* \right\|_2^2$ from both sides, noting that $X_j - Z\pi_j^* = Z\pi_j^* + \eta_j - Z\pi_j^* = \eta_j$ and

rearranging, we find

$$\frac{1}{n}\eta_j^t Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right) + \lambda\left(\left\|\pi_j^*\right\|_1 - \left\|\hat{\pi}_j^{\text{LASSO}}\right\|_1\right) \geqslant \frac{1}{2n}\left\|Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right)\right\|_2^2.$$

Using Hölder's inequality to further bound the left hand side of from above, we find (2.50). □

Lemma 48. If

$$\lambda \geqslant \frac{2}{n}\left\|Z^t\eta_j\right\|_\infty$$

then

$$\frac{1}{n}\left\|Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right)\right\|_2^2 \leqslant 12\lambda\left\|\pi_j^*\right\|_1.$$

Proof. Let's start with (2.50). Note that the lower bound $\frac{1}{2n}\left\|Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right)\right\|_2^2$ can be further bounded from below by zero. Then we have with the triangle inequality

$$\begin{aligned}
0 &\leqslant \left\|\frac{1}{n}\eta_j^t Z\right\|_\infty \left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 + \lambda\left(\left\|\pi_j^*\right\|_1 - \left\|\hat{\pi}_j^{\text{LASSO}}\right\|_1\right) \\
&= \left(\left\|\frac{1}{n}\eta_j^t Z\right\|_\infty - \lambda\right)\left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 + \lambda\left(\left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 + \left\|\pi_j^*\right\|_1 - \left\|\hat{\pi}_j^{\text{LASSO}}\right\|_1\right) \\
&\leqslant \left(\left\|\frac{1}{n}\eta_j^t Z\right\|_\infty - \lambda\right)\left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 + \lambda\left(\left\|\hat{\pi}_j^{\text{LASSO}}\right\|_1 + \left\|\pi_j^*\right\|_1 + \left\|\pi_j^*\right\|_1 - \left\|\hat{\pi}_j^{\text{LASSO}}\right\|_1\right) \\
&= \left(\left\|\frac{1}{n}\eta_j^t Z\right\|_\infty - \lambda\right)\left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 + 2\lambda\left\|\pi_j^*\right\|_1 \\
&\leqslant -\frac{1}{2}\lambda\left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 + 2\lambda\left\|\pi_j^*\right\|_1 = \frac{\lambda}{2}\left(4\left\|\pi_j^*\right\|_1 - \left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1\right).
\end{aligned}$$

Comparing the last estimate with the lower bound 0, we find

$$\left\|\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right\|_1 \leqslant 4\left\|\pi_j^*\right\|_1 \tag{2.51}$$

Now let's consider (2.50) again:

$$\begin{aligned}
\frac{1}{2n}\left\|Z\left(\hat{\pi}_j^{\text{LASSO}}-\pi_j^*\right)\right\|_2^2 &\leqslant \left\|\frac{1}{n}\eta_j^t Z\right\|_\infty \left\|\hat{\pi}_j^{\text{LASSO}}-\pi_j^*\right\|_1+\lambda\left(\left\|\pi_j^*\right\|_1-\left\|\hat{\pi}_j^{\text{LASSO}}\right\|_1\right)\\
&\leqslant \left\|\frac{1}{n}\eta_j^t Z\right\|_\infty \left\|\hat{\pi}_j^{\text{LASSO}}-\pi_j^*\right\|_1+\lambda\left(\left\|\pi_j^*-\hat{\pi}_j^{\text{LASSO}}\right\|_1\right) \qquad \text{(triangle ineq)}\\
&\leqslant \frac{1}{2}\lambda\left\|\hat{\pi}_j^{\text{LASSO}}-\pi_j^*\right\|_1+\lambda\left(\left\|\pi_j^*-\hat{\pi}_j^{\text{LASSO}}\right\|_1\right)\\
&= \frac{3}{2}\lambda\left\|\hat{\pi}_j^{\text{LASSO}}-\pi_j^*\right\|_1\\
&\leqslant \frac{3}{2}\lambda 4\left\|\pi_j^*\right\|_1=6\lambda\left\|\pi_j^*\right\|_1.
\end{aligned}$$

Multiplying by 2 gives the desired result. □

So far, all arguments were done without using the distribution of Z or η_j. This came at a price: the statement in lemma 48 is conditional on the inequality

$$\lambda \geqslant \frac{2}{n}\left\|Z^t\eta_j\right\|_\infty$$

which, depending on the realizations of Z and η_j, may or may not hold. Of course, we are interested in choosing λ large enough to ensure that the event holds with "large" probability. For this, we use our Assumptions 3 and 4.

Lemma 49. Suppose Assumption 3 and 4 are satisfied. Set $t \geqslant 0$ and choose $\mathfrak{J} \subset \{1,\ldots,p\}$ arbitrarily. Then with probability at least

$$1-2\exp\left(\log(d\,|\mathfrak{J}|)-c_B\min\left\{\frac{t^2}{\rho_Z^2\rho_\eta^2},\frac{t}{\rho_Z\rho_\eta}\right\}n\right)$$

we have

$$\max_{j\in\mathfrak{J}}\frac{1}{n}\left\|Z^t\eta_j\right\|_\infty \leqslant t.$$

Proof. We have

$$\frac{1}{n}\left\|Z^t\eta_j\right\|_\infty = \max_{k\in\{1,\dots,d\}}\left|\frac{1}{n}\sum_{i=1}^n Z_{i,k}\eta_{i,j}\right|.$$

We note that $Z_{i,k}\eta_{i,j}$ is independent over i by Assumption 3. Also by Assumption 3, the expectation of $Z_{i,k}\eta_{i,j}$ is zero. In Assumption 4, we have assumed that $Z_{i,k}$ are sub-Gaussian with sub-Gaussian norm at most ρ_z and that $\eta_{i,j}$ is sub-Gaussian with sub-Gaussian norm at most ρ_η. We know that the product of two sub-Gaussians is sub-exponential with the sub-exponential norm bounded by the product of the sub-Gaussian norms. Hence we can apply Bernstein's inequality to infer that for any $j \in \{1,\dots,p\}$ we have for any $t \geqslant 0$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n Z_{i,k}\eta_{i,j}\right| \geqslant t\right] \leqslant 2\exp\left(-c_B\min\left\{\frac{t^2}{\rho_Z^2\rho_\eta^2}, \frac{t}{\rho_Z\rho_\eta}\right\}n\right)$$

Using a union bound, we find for any $t \geqslant 0$

$$\begin{aligned}
\mathbb{P}\left[\max_{j\in\mathfrak{J}}\max_{k\in\{1,\dots,d\}}\left|\frac{1}{n}\sum_{i=1}^n Z_{i,k}\eta_{i,j}\right| \geqslant t\right] =& \mathbb{P}\left[\bigcup_{\substack{j\in\mathfrak{J}\\k\in\{1,\dots,d\}}}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_{i,k}\eta_{i,j}\right| \geqslant t\right\}\right] \\
\leqslant& \sum_{\substack{j\in\mathfrak{J}\\k\in\{1,\dots,d\}}}\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n Z_{i,k}\eta_{i,j}\right| \geqslant t\right] \\
\leqslant& 2d\left|\mathfrak{J}\right|\exp\left(-c_B\min\left\{\frac{t^2}{\rho_Z^2\rho_\eta^2}, \frac{t}{\rho_Z\rho_\eta}\right\}n\right) \\
=& 2\exp\left(\log(d\left|\mathfrak{J}\right|) - c_B\min\left\{\frac{t^2}{\rho_Z^2\rho_\eta^2}, \frac{t}{\rho_Z\rho_\eta}\right\}n\right).
\end{aligned}$$

This completes the proof. □

We now combine the results we have shown so far in the following Corollary.

Corollary 50. Suppose Assumption 3 and 4 are satisfied. Choose some $\mathfrak{J} \subset \{1,\dots,p\}$. Fix some

$0 < \tau < \sqrt{\frac{n}{\log(d|\mathfrak{J}|)}}$ and set

$$\lambda = 2\rho_z \rho_\eta \sqrt{\frac{\log(d\,|\mathfrak{J}|)}{n}} \tau$$

Then with probability at least

$$1 - 2\exp\left(\log(d\,|\mathfrak{J}|)\left(1 - c_B\tau^2\right)\right) \tag{2.52}$$

it holds for all $j \in \mathfrak{J}$ that

$$\frac{1}{n}\left\|Z\left(\hat{\pi}_j^{\text{LASSO}} - \pi_j^*\right)\right\|_2^2 \leqslant 24\rho_z\rho_\eta\sqrt{\frac{\log(d\,|\mathfrak{J}|)}{n}}\tau\left\|\pi_j^*\right\|_1. \tag{2.53}$$

Proof. Combine Lemma 49 and Lemma 48. □

In particular, if n, d and $|\mathfrak{J}|$ tend to infinity such that $\frac{\log(d|\mathfrak{J}|)}{n}$ tends to zero, we can choose $\tau = \frac{1}{2\sqrt{c_B}}$ and see that with probability converging to one, the LASSO prediction error is bounded by a constant times $\sqrt{\frac{\log(d|\mathfrak{J}|)}{n}}$. This is the so-called slow rate of the LASSO.

Chapter 3

Demand Estimation with Finitely Many Consumers

Coauthored with Thomas Wiemann*

3.1 Introduction

The problem of estimating demand parameters in settings with endogenous prices arises frequently in empirical economics. A key assumption in commonly applied demand models is that population-level market shares are observed without error. In practice, however, market shares are often estimated as averages of consumer choices. When the number of consumers over which these averages are computed is finite, these shares will be estimated with error. Estimation errors in the market shares propagate to estimation errors in the demand parameters which do not "average out" due to the nonlinearity of the discrete choice model.

One symptom of this is the so-called "zero-market-share problem" which arises when products are not purchased in every market (e.g., Dubé et al., 2021). Zero-valued market shares imply that

*Kenneth C. Griffin Department of Economics, University of Chicago

the conventional demand estimators are not defined and cannot be computed, making the zero-market-share problem particularly salient. In response to zero-valued market shares, researchers often use ad-hoc solutions such as removing market-product combinations with no purchases. As noted in Dubé et al. (2021), researchers may not always be explicitly aware of ad-hoc manipulation of the data. Common datasets such as retail scanner data from IRI and Nielsen, for example, only report products with positive purchases. These data manipulations may however introduce additional biases and render existing theoretical guarantees on conventional estimators inapplicable.

In this paper, we propose a new estimator of demand parameters suitable for settings with estimated and possibly zero-valued market shares. The small but important departure from the demand models of Berry (1994) and Berry et al. (1995) is that we consider the observed market shares to be generated by a finite number of consumers. The estimator is based on a constrained optimization problem constructed by generalizing the mathematical program with equilibrium constraints (MPEC) formulation of Dubé et al. (2012) using known bounds on the estimation error in the observed market shares. We dub the estimator Estimated/Zero-share MPEC (EZ-MPEC) to highlight the applicability of the estimator in settings with estimated and zero-valued market shares. Our theoretical results show consistency of the estimator as the number of markets (T) and the number of consumers in each market (n) grow such that $\log(T)/n \to 0$. We further provide confidence intervals via test inversion under the same regime.

Although we focus on demand estimation based on random coefficient logit models, our results generalize to a larger class of demand estimators, including nonparametric demand estimation as in Tebaldi et al. (2019). To the best of our knowledge, this is the first application of finite-sample concentration bounds to the construction of a demand estimator for settings with estimated market shares and endogenous prices. In simulations, we highlight prevalence of biases arising through estimated market shares and ad-hoc data manipulations as well as illustrate the good performance of the EZ-MPEC estimator.

This paper contributes to the growing literature aiming to resolve the zero-market-share prob-

lem. One strand of this literature views the occurrence of zeros in the observed market shares as a rejection of the conventional random coefficient demand model and proposes extensions that motivate population-level market shares of zero. Gandhi et al. (2020) propose an asymptotic regime in which products are either "safe" with a population-level market share bounded away from zero or "risky" with a population-level market share of zero, and Dubé et al. (2021) incorporate consideration sets of consumers. The other strand of literature, to which we contribute primarily, focuses on sampling error of the market shares – and associated positive probability of zero-valued market shares – when these are constructed based on finitely many consumer purchases. To characterize the finite sample uncertainty in the observed market shares, Hortaçsu et al. (2021) consider a Bayesian IV approach and assume consumer arrivals follow a Poisson process. Regardless of the cause of the zero-shares considered by these approaches, however, the estimators proposed in existing literature rely on additional data not conventionally needed for demand estimation. In particular, instruments informative about the identity of safe products, the consumers' consideration sets, or the consumers' search behavior is needed, respectively. In contrast, the estimator proposed in this paper does not make a substantial structural deviation of the popular demand model and requires no additional data beyond the number of consumers in every market.

We also contribute to the literature on demand estimation with estimated market shares. Berry et al. (2004) and Freyberger (2015) develop asymptotic distributions of the random coefficient logit demand estimator with estimated market shares when the number of products J or the number of markets T grow, respectively, and show that the number of customers n must grow sufficiently quickly to achieve asymptotic normality. Freyberger (2015) further shows that the conventional estimator is not centered at the true value unless $\sqrt{T}/n \to 0$. The bias correction and covariance adjustments suggested by this literature are not applicable, however, when an estimated market share is zero-valued. In contrast, the confidence intervals we provide hold for $\log(T)/n \to 0$ and can be computed regardless of whether zero-valued market shares occur in the data. This appears particularly important as the settings with small numbers of consumers in which the finite-sample

corrections of Freyberger (2015) are most likely to be relevant are also the settings in which they are least likely to be applicable due to increased probability of zero-valued market shares in the data.

The rest of the paper proceeds as follows: Section 3.2 reviews the random coefficient logit model and discusses consequences of estimated market shares. Section 3.3 presents and discusses the EZ-MPEC estimator. Section 3.5 provides Monte Carlo simulations to illustrate finite sample performance of the proposed estimator to conventional alternatives. Section 3.6 concludes.

3.2 The random coefficient logit model

This section briefly reviews the random coefficient demand model of Berry (1994) and Berry et al. (1995), discusses the finite number of consumers adaptation, and illustrates its consequences for estimation.

We begin by defining the demand model in a single market. Consider a consumer i who chooses an alternative from $\mathcal{J} \equiv \{0, \ldots, J\}$ that maximizes their utility

$$Y_i = \underset{j \in \mathcal{J}}{\arg\max} \; X_j^\top \beta_i + \xi_j + \varepsilon_{i,j}, \tag{3.1}$$

where X_j are observed product characteristics of the jth alternative including prices, β_i is a consumer-specific parameter vector β_i, ξ_j is the corresponding unobserved demand shock, and ε_{ij} is a consumer and product-specific latent utility shock. Throughout, $j = 0$ is the outside option with utility normalized to zero. Let X be the $J \times d_X$ vector with X_j as elements and define ξ analogously. To obtain the random coefficient logit model, we place distributional assumptions on the demand coefficients and latent utility shocks.

Assumption 8. Consumers choose an alternative from $\mathcal{J} \equiv \{0, \ldots, J\}$ via (3.1) $\forall i = 1, \ldots, n$. The latent utility shocks $\varepsilon_{i,j}$ are i.i.d. T1EV. Customer preference parameters β_i are i.i.d. multivari-

ate normal with parameter $\theta \equiv (\mu, \Sigma)$.

Integrating over ε_i and β_i, results in the common expression for conditional choice probabilities (CCPs):

$$\pi_j(X, \xi; \theta) \equiv \Pr(Y_i = j | X, \xi; \theta) = \int \frac{\exp\left(X_j\beta + \xi_j\right)}{1 + \sum_{k=1}^{J} \exp\left(X_k\beta + \xi_k\right)} dF(\beta; \theta). \tag{3.2}$$

Since the latent demand shocks ξ are unobserved by the econometrician but potentially correlated with observed product characteristics such as prices, researchers frequently leverage instrumental variables Z for estimation of the demand parameters. Assumption 9 states a moment condition frequently employed in practice.

Assumption 9. There exists an instrument Z such that $E\left[Z^\top \xi\right] = 0$.

To allow for the application of common linear IV methodology despite the non-linear dependence of the CCPs in (3.2), Berry (1994) shows that for any (X, θ), there exists a bijective map between the value of the CCPs and the latent demand shocks. In particular, for all $s \in (0,1)^J : \|s\|_1 < 1$, there exists a unique $\xi \in \mathbb{R}^J$ such that $\pi(X, \xi; \theta) = s$, where $\pi(X, \xi; \theta)$ denotes the $J \times 1$ vector of CCPs. Replacing the unobserved demand shocks in the moment condition of Assumption 9 with these solutions $\xi(X, s; \theta)$, then motivates GMM estimators that replaces s with the strictly positive CCPs.

In practice the econometrician may not observe market shares sampled directly from the CCPs as postulated in Berry et al. (1995). Instead, market shares are frequently estimated as sample averages over consumer choices. Assumption 10 highlights this small but critical deviation from the conventional demand models with observed CCPs.

Assumption 10. Observed market shares are sample averages of consumer choices

$$\hat{S}_j^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i = j\}, \ \forall j \in \mathcal{J}.$$

Settings with observed market shares constructed from a finite sample of consumer purchase decisions as considered here introduce several challenges to demand estimation. In particular, due to the nonlinear nature of the conventional demand model in equation (3.2), the sampling error in the market shares does not straightforwardly average out and instead introduces an incidental parameter problem. For example, consider the conventional GMM estimator in a setting with many markets given by

$$\hat{\theta}_{blp} = \arg\min_{\theta \in \Theta} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t^\top \xi(X_t, S_t; \theta) \right)^\top W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t^\top \xi(X_t, S_t; \theta) \right), \tag{3.3}$$

where is $(S_t, X_t, Z_t) \overset{iid}{\sim} (\pi(X, \xi; \theta_0)), X, Z)$ is a sample of $t = 1, \ldots, T$ markets, and W_T is a positive-definite weighting matrix, often chosen to be $W = (\sum_{t=1}^{T} Z_t Z_t^\top)^{-1}$. Freyberger (2015) characterizes the asymptotic distribution of $\hat{\theta}_{blp}$ when S_t are replaced with estimated market shares $\hat{S}_t^n$ as $T \to \infty$. Similarly, Berry et al. (2004) characterizes the analogue estimator with estimated market shares in a setting with many products.[1] The authors show that unless the number of consumers grow at sufficiently fast rate, the conventional estimator is not $\sqrt{T}$ or $\sqrt{J}$ Gaussian. Freyberger (2015) further shows that the asymptotic distribution of the estimator as the number of markets grow is not centered at the true value unless $\sqrt{T}/n \to c$ for some finite constant c. The author proposes a bias correction to improve finite sample performance, which is shown to improve performance in Monte Carlo simulations.

In addition to the incidental parameter problem, estimated market shares can be cause for the zero-market-share problem. This is because for any product-market combination, $n\hat{S}_{jt}^{(n)} \sim$ Binomial$(\pi_j(X_t, \xi_t; \theta_0), n)$, implying a strictly positive probability for the event that some products do not have purchases in every market (i.e., $\Pr(\exists j, t : \hat{S}_{jt} = 0) > 0$). The occurrence of zero-

[1] Berry et al. (2004) and Freyberger (2015) also accommodate for sampling uncertainty from the Monte Carlo integration of the integral in equation (3.2). We focus on sampling uncertainty from the estimation of market shares only here, but note that our application of concentration inequalities straightforwardly extends to the Monte Carlo integration error.

valued market shares in particular has received increased attention in recent literature due to both the abundance of economic settings with no observed purchases for some products and its severe consequences for conventional estimation approaches. An important aspect of estimators such as $\hat{\theta}_{blp}$ leveraging the demand inversion directly is that they require the set of market shares $(S_t)_{t=1}^T$ to be positive since otherwise $\xi(\cdot)$ is not defined.

The infeasiblity of $\hat{\theta}_{blp}$ with estimated zero-valued market shares is particularly evident using the MPEC formulation proposed by Dubé et al. (2012):

$$\begin{aligned} \min_{(\theta,\xi)} \quad & \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right)^\top W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right) \\ \text{s.t.} \quad & S_{jt} = \pi_j(X_t, \xi_t; \theta),\ \forall (j,t) \in \mathcal{J} \times \mathcal{T}, \end{aligned} \tag{3.4}$$

where $\mathcal{T} \equiv \{1, \ldots, T\}$ is the set of observed markets. Dubé et al. (2012) show that MPEC is equivalent to the previously defined GMM estimator $\hat{\theta}_{blp}$. Since the domain of the CCPs $\pi_j(\cdot)$ is strictly positive, presence of any vector of market shares S_t that is not strictly positive implies that no feasible solution to the mathematical program exists. Replacing $(S_t)_{t=1}^T$ with their estimated counterparts can thus result in infeasibility of $\hat{\theta}_{blp}$. Further, the bias corrections suggested by Freyberger (2015) cannot be computed for the same reasons.

In practice, researchers often apply ad-hoc manipulations to their data when zero-valued market shares arise to be able to apply conventional random coefficient logit estimators. Popular approaches appear to be: 1) Replacing the zero-valued market shares with an arbitrary small number (e.g., $0.5/n$), or 2) removing the product-market combinations from the sample that are associated with zero-valued market shares (Quan and Williams, 2018; Gandhi et al., 2020; Dubé et al., 2021).[2] The bias of the first approach is likely sensitive to the specific small number used (see, e.g., Dubé

[2]Alternatively, researchers may choose to aggregate purchases across products and markets until estimated shares are strictly positive. This introduces measurement error and limits the type of question that can be answered as products and markets are combination are often artificial, smoothing across relevant heterogeneity and making the results challenging to interpret. Quan and Williams (2018) give an augmented nested logit model under which local demand heterogeneity is identified using aggregated rather than local market shares.

et al., 2012). We are more concerned with the second approach of truncating the data, as this can lead to substantial selection bias. In particular, while by assumption $E[Z_t^\top \xi_t] = 0$ for all markets for identification, the moment conditions $E[Z_t^\top \xi_t | \hat{S}_t^{(n)} > 0]$ do not hold in general because conditional on $\hat{S}_t^{(n)} > 0$ the latent demand shocks are less likely to be negative. The next section develops an alternative estimator that explicitly takes sampling error in the marketing shares into account and can be applied to settings with zero-valued market shares.

3.3 Estimation

To address the issues arising from estimation errors in market-shares, we combine the MPEC estimator (3.4) developed by Dubé et al. (2012) with finite-sample confidence intervals on the observed market shares. Throughout, we consider an i.i.d. sample across markets.

Assumption 11. The data is an i.i.d. sample $\{(\hat{S}_t^{(n)}, X_t, Z_t)\}_{t=1}^T$ from $(\hat{S}^{(n)}, X, Z)$.

We propose to consider any solution to

$$\begin{aligned} \min_{(\theta,\xi)} \quad & \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right)^\top W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right) \\ \text{s.t.} \quad & \hat{S}_{jt}^{(n)} \in C_{n,T}^j(X_t, \xi_t; \theta, \alpha),\ \forall (j,t) \in \mathcal{J} \times \mathcal{T}. \end{aligned} \tag{3.5}$$

Compared to (3.4), we replace the constraint that observed market shares equal model-implied expected CCPs with the constraint that observed market shares must be contained in a set $C_{n,T}$. As in the equality constraints of (3.4), $C_{n,T}$ depends on product characteristics, the latent demand shocks, and the demand parameter. In addition, the set depends on the number of consumers n, the number of markets T, and a hyper-parameter $\alpha \in (0,1)$.[3]

The purpose of the sets $C_{n,T}$ is to bound the sampling errors in the estimation of the market shares. We choose this set to be a joint confidence set that covers the true population-level market

[3] $C_{n,T}$ further depends on the number of products J. Since J is fixed throughout, we omit this dependence.

shares with probability at least $1 - \alpha$ for any finite number of consumers. These finite-sample confidence intervals provide a probabilistic bound on the deviations of the estimated market shares from their population-values. These joint confidence intervals are constructed in two steps: First, we consider a particular product-market combination and derive a probabilistic bound on the estimation error in the observed market share. Second, we adjust the marginal confidence level to achieve uniform coverage at a desired rate. We now discuss both steps in turn.

Multiple methods to derive finite-sample confidence sets exist. In choosing a method, the researcher faces a trade-off between tractability and tightness. On one end of the spectrum are confidence-intervals based on Hoeffding's inequality. These confidence intervals lend themselves to straightforward computation of the EZ-MPEC estimator as they correspond to constraints which are linear in the model-implied choice probabilities. EZ-MPEC based on these bounds thus exhibits the same Jacobian and Hessian used for solving infeasible MPEC.[4] However, Hoeffding's inequality is conservative, allowing for larger sampling error in the market shares than necessary for the desired nominal coverage level. On the other end of the spectrum, binomial quantiles have exact coverage but imply an increased computational burden on the constrained optimization problem due to the non-linear dependence of the confidence sets on the model-implied CCPs.[5]

Given a choice of bound, the marginal confidence intervals need to be adjusted to achieve joint coverage for all estimated market shares. We consider simple adjustments based on the union bound (Bonferroni) in Proposition 51. These adjustments are conservative as they allow for arbitrary dependence across products without exploiting the structure of the random coefficient logit model. Sharp bounds on the sampling error may be obtained using multinomial quantiles. Unfortunately, multinomial quantiles are computationally intractable for even small numbers of products. We thus focus on bounds that are more likely to be useful in practice.

[4]We note that $C_{n,T}$ based on Hoeffding's inequality also allows using finite-sample confidence intervals for the nonparametric demand model proposed in Tebaldi et al. (2019) which requires linearity for tractability. See Appendix 3.8.4 for further illustration.

[5]In Appendix 3.8.5, we give implementation details for feasible MPEC using the binomial quantiles.

Proposition 51. Let assumptions 8-11 hold. Fix $\alpha \in (0,1)$. Then with Hoeffding's inequality, it holds that

$$Pr\left(\exists(j,t) \in \mathcal{J} \times \mathcal{T} : \ |\hat{S}_{jt} - \pi_j(X_t, \xi_t; \theta_0)| \geqslant \sqrt{\frac{\log\left(\frac{2J}{1-\sqrt[T]{1-\alpha}}\right)}{2n}}\right) \leqslant \alpha,$$

$\forall n \in \mathbb{N}_{++}$. With binomial quantiles, it holds that

$$Pr\Bigg(\exists(j,t) \in \mathcal{J} \times \mathcal{T} : \ \hat{S}_{jt} \notin \Bigg[\frac{1}{n} F^{-1}_{\text{Bin}}\left(\frac{1-\sqrt[T]{1-\alpha}}{J}, \pi_j(X_t, \xi_t; \theta_0), n\right), \\ \frac{1}{n} F^{-1}_{\text{Bin}}\left(1 - \frac{1-\sqrt[T]{1-\alpha}}{J}, \pi_j(X_t, \xi_t; \theta_0), n\right)\Bigg]\Bigg) \leqslant \alpha,$$

$\forall n \in \mathbb{N}_{++}$, where $F^{-1}_{\text{Bin}}(\cdot, p, n)$ denotes the quantile function of the binomial distribution with n trials and success probability $p \in (0,1)$. For $J = 1$, the second inequality holds with equality.

Proposition 51 provides explicit bounds on the estimation error of all market shares simultaneously that hold with probability at least $1 - \alpha$. Since these bounds depend on the true CCPs, $\pi(X_t, \xi_t; \theta_0)$, they can be leveraged in estimation. For example, when using bounds based on Hoeffding's inequality, the EZ-MPEC estimator is given by

$$\begin{aligned} \min_{(\theta,\xi)} \quad & \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right)^\top W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right) \\ \text{s.t.} \quad & \hat{S}^{(n)}_{jt} - \pi_j(X_t, \xi_t; \theta) \leqslant \sqrt{\frac{\log\left(\frac{2J}{1-\sqrt[T]{1-\alpha}}\right)}{2n}}, \quad \forall j,t, \\ & \pi_j(X_t, \xi_t; \theta) - \hat{S}^{(n)}_{jt} \leqslant \sqrt{\frac{\log\left(\frac{2J}{1-\sqrt[T]{1-\alpha}}\right)}{2n}}, \quad \forall (j,t) \in \mathcal{J} \times \mathcal{T}. \end{aligned}$$

Remark 52. Although Proposition 51 considers the random coefficient logit model due to its popularity in demand estimation, we highlight that these bounds on the market share sampling

error do not depend on the specific demand model. In particular, researchers may substitute CCPs implied by any discrete choice model for $\pi(\cdot)$ that admits population-level demand inversion.

3.4 Asymptotic Properties

This section provides formal consistency and inference results based on large n and T asymptotics. We begin with listing additional assumptions. [6]

Assumption 12. $\exists \gamma \in (0,1)$ such that $P(\pi(X,\xi;\theta_0) \in [\gamma, 1-\gamma]^J) = 1$.

Assumption 13. Θ, $\operatorname{supp} X$, and $\operatorname{supp} Z$ are compact.

Assumption 14. The matrix $\frac{1}{T}\sum_{t=1}^{T} Z_t^\top Z_t$ has full rank and is stochastically bounded, i.e., $\forall \varepsilon > 0$ there exists an $M(\varepsilon)$ such that $\Pr\left(\|\frac{1}{T}\sum_{t=1}^{T} Z_t^\top Z_t\| > M(\varepsilon)\right) < \varepsilon$.

Assumption 15. $\forall \delta > 0,\ \exists M(\delta) > 0$, such that

$$\lim_{T\to\infty} \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} \|G_T(\theta) - G_T(\theta_0)\| \geqslant M(\delta)\right) = 1,$$

where G_T is the objective function in (3.3).

Assumptions 12 and 13 restrict the support of the population-level market shares, the product characteristics, and the parameter space. Assumption 14 places moment restrictions on the instrument vectors. Finally, Assumption 15 assumes identification of the demand parameter from the moment conditions in Assumption 9. These assumptions are analogous to those assumed for asymptotic analysis in Freyberger (2015).

The assumptions are sufficient to show convergence of the EZ-MPEC estimator to the demand parameters θ_0 for many markets T and many consumers per market n at the rate $\log(T)/n \to 0$. For asymptotic analysis, we let the confidence parameter α depend on (n,T).

[6] Notation: When A is a matrix, $\|A\| = \text{trace}(A^tA)^{1/2}$. Else, $\|\cdot\|$ denotes the Euclidean norm. Further, we denote a neighborhood by $\mathcal{N}_{x_0}(\delta) \equiv \{x \in \mathcal{X}_0 : \|x - x_0\| \leqslant \delta\}$.

Theorem 53. Let assumptions 8 to 15 hold. If in addition $\alpha_{n,T} \in (0,1) : \alpha_{n,T} = o_p(1)$ and $\log(T) = o_p(n)$, then $\forall \epsilon > 0$,

$$\lim_{n,T \to \infty} \Pr\left(\sup_{\tilde{\theta} \in \Theta^*_{n,T}} \left\| \tilde{\theta} - \theta_0 \right\| > \epsilon \right) = 0,$$

where $\Theta^*_{n,T}$ denotes the $\arg\min$ of the EZ-MPEC estimator in (3.5) with bounds based on Proposition 51 and hyperparameter $\alpha_{n,T}$, and θ_0 are the true demand parameters.

Remark 54. While Theorem 53 is formulated for EZ-MPEC based on the inequalities of Proposition 51, it also applies to many other methods to derive finite-sample confidence intervals. (3.5) combined with any method to construct finite-sample confidence intervals enjoys the properties of Theorem 53 as long as the confidence intervals are not larger than the confidence intervals based on Hoeffding's inequality. This includes, in particular, binomial and multinomial quantiles.

To obtain confidence intervals with sufficient coverage of θ_0, we rely on inversion of tests of null hypotheses of the form

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

For this purpose, we propose the test statistic $\hat{G}_T$ given by

$$\begin{aligned} \hat{G}_T(\alpha) \equiv \min_{\{\tilde{\xi}_t\}_{t=1}^T} \quad & T \left(\frac{1}{T} \sum_{t=1}^T Z_t^\top \tilde{\xi}_t \right)^\top \left(\frac{1}{T} \sum_{t=1}^T \hat{\xi}_t^\top Z_t Z_t^\top \hat{\xi}_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T Z_t^\top \tilde{\xi}_t \right) \\ \text{s.t.} \quad & \hat{S}_{jt}^{(n)} \in C^j_{n,T}(X_t, \tilde{\xi}_t; \theta, \alpha), \ \forall (j,t) \in \mathcal{J} \times \mathcal{T}, \end{aligned} \tag{3.6}$$

where $(\hat{\xi}_t)_{t=1}^T$ are consistent estimates of the latent demand shocks obtained from computing EZ-MPEC in a first step. The test statistic is akin to the χ^2 test statistic conventionally leveraged in GMM inference of the random coefficient logit models of Berry et al. (1995), but accounts for the sampling error in the market shares. Theorem 55 gives a test based on (3.6) with controlled size.

Theorem 55. If the assumptions of Theorem 53 hold, in particular, $\log(T) = o_p(n)$ and $\alpha_{n,T} \in (0,1) : \alpha_{n,T} = o_p(1)$, then under H_0

$$\limsup_{n,T\to\infty} E\left[\mathbb{1}\{\hat{G}_T(\alpha_{n,T}) > c_K^{1-\tau}\}\right] \leqslant \tau,$$

for any $\tau \in (0,1)$, where $c_K^{\alpha/2}$ is the $1-\frac{\alpha}{2}$ quantile of a χ^2 distribution with K degrees of freedom.

Importantly, Theorem 53 and Theorem 55 hold as $\log(T) = o_p(n)$. In contrast, Freyberger (2015) shows that the naive estimator that replaces the population-level market shares in infeasible MPEC with their estimates is asymptotically normal only $\sqrt{T} = o_p(n)$. As $\log(T) << \sqrt{T}$ for large T, our estimator is robust to situations when there are much fewer consumers per market. Even stronger, Freyberger (2015) shows that in the asymptotic regime we consider, the naive estimator that replaces the population-level market shares in infeasible MPEC with their estimates incurs a bias that is not bounded in probability when rescaled with $\sqrt{T}$.

3.5 Monte Carlo Simulation

We conduct Monte Carlo simulations to illustrate the implications of estimated market shares for conventional random coefficient logit estimators that remove zero-valued market shares from their data and highlight improvements of the proposed EZ-MPEC estimator based on binomial quantiles. The setting is similar to the simulations reported in Dubé et al. (2012).

Each of the simulations considers a setting with $J = 5$ products (and an outside option) and $T = 50$ markets. We let the number of consumers per market vary across simulations to analyze settings with varying sampling uncertainty in the estimated market shares. Each consumer makes their purchasing decision with CCP associated with product j given in Assumption 8, where

$$X_j^\top \beta_i = \beta_i^0 + W_j^t \beta_i^w - p_j \beta_i^p,$$

and $\beta_i = (\beta_i^0, (\beta_i^w)^t, \beta_i^p)$. Throughout, we take

$$W_j = \begin{bmatrix} W_j^1 \\ W_j^2 \\ W_j^3 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.55 & -.25 & 0.2 \\ -.25 & 1.55 & 0.2 \\ 0.2 & 0.2 & 1.55 \end{bmatrix} \right), \quad \text{and} \quad \xi_j \sim N(0,1),$$

and let the endogenous price be given by $p_j = \max(0.1\xi_j + e_j, , 0.01)$, with $e_j \sim N(0,1)$. Note that endogeneity is introduced through dependence of price on the latent demand shock ξ_j. For each product, there is an additional underlying instrument vector Z_j of dimension 6×1 with entries generated $Z_j^m = U(0,1) + 0.25e_j$, where $U(0,1)$ is the realization of a standard uniformly distributed random variable. In estimation, we construct a higher order standard polynomial expansion of Z_j with the product characteristics W_j.[7]

Our analysis applies the proposed EZ-MPEC estimator in (3.5) where the bounds C_n^α are based on the Binomial quantiles. We choose $\alpha = 0.1$ so that the set of all $J \times T$ sampling error conditions hold jointly with probability at least 0.9.

We begin by considering a data generating process where the demand parameters β_i are fixed across consumers at $\beta^0 = 0$, $\beta^w = (1\ 1\ \text{-}1)^\top$, and $\beta^p = -3$. In this logit setting without random coefficients, demand inversion with population-level market shares motivates a simple two-stage least squares (TSLS) estimator with $\log(S_j/S_0)$ as the second stage left-hand side variable. We compare EZ-MPEC to both the infeasible TSLS estimator using the population-level market shares as well as to the ad-hoc TSLS estimator that removes product-market combinations with zero-valued market shares.

Table 3.1 presents the results for 1,000 simulations. We focus on the median absolute error (MAE) of the demand parameter corresponding to the endogenous price variable ($\beta^p = -3$) as the evaluation criterion in columns (1)-(3) to assess both centrality and dispersion. Column (4) gives

[7] Following a similar approach as in Dubé et al. (2012), we consider $Z_j, Z_j^2, Z_j^3, W_j, W_j^2, W_j^3, \prod_{m=1}^6 Z_j^m, \prod_{k=1}^3 W_j^k, Z_j \cdot W_j^1, Z_j \cdot W_j^2$. This results in a total of 42 moment conditions.

the average share of zero-valued market shares in the sample, which corresponds to the share of the $J \times T$ observations removed from the data in column (2).

TSLS with population-level market shares in column (1) of Table 3.1 has an MAE of approximately 0.09.[8] In contrast, the TSLS estimator with estimated sample shares (TSLS: Ad-Hoc) has substantially higher MAE as large as 0.44 at 250 consumers per market. This difference reduces as the number of consumers per market grow and corresponding share of zero-valued market shares decreases, yet, even with 5000 consumers per market do not suffice to estimate the market shares at sufficient accuracy to ignore their sampling error.

The EZ-MPEC estimator in column (3) of Table 3.1 improves over the TSLS estimator with estimated market shares. In particular for markets with a small numbers of consumers between 500-1000, the differences are substantial. For example, at 750 consumers per market, the EZ-MPEC estimator as an MAE that is 0.103 smaller than TSLS: Ad-Hoc. Given that the sampling uncertainty of the infeasible TSLS estimator with population-level market shares corresponds to an MAE of approximately 0.09, this highlights that ignoring sampling error in the market shares can lead to qualitatively very different results. As the number of consumers increase and the share of zero-valued market shares decreases, the incidental parameter and selection bias of the TSLS estimator with estimated market shares decrease, reducing the magnitude of the MAE to that of the EZ-MPEC estimator as expected.

[8]Note that because consumer markets have no implications for the population-level market shares, any differences in the corresponding TSLS estimator are due to sampling uncertainty of the T markets.

Table 3.1: Mean Absolute Error for DGP without Random Coefficients

# Consumers (zero-share)	TSLS: Infeasible	TSLS: Ad-Hoc	EZ-MPEC
	(1)	(2)	(3)
500 (0.133)	0.094	0.317	0.231
750 (0.110)	0.091	0.289	0.186
1000 (0.095)	0.088	0.260	0.185
2000 (0.068)	0.091	0.191	0.166
3000 (0.054)	0.084	0.167	0.154
4000 (0.046)	0.088	0.173	0.150
5000 (0.040)	0.092	0.150	0.158

Notes. Results based on 1,000 Monte Carlo simulations. TSLS: Infeasible and TSLS: Ad-Hoc denote two-stage least squares estimators using the population-level and estimated market shares, respectively. TSLS: Ad-Hoc is computed using only those product-market combinations with positive estimated market shares. Parentheses state the average fraction of observations with zero-valued estimated market shares.

In a second set of simulations, we consider a data-generating process based on the random coefficient logit model. In particular, the consumer-specific demand parameters β_i are uncorrelated and generated as $\beta_i^0 \sim N(0, 0.25)$, $\beta_i^w \sim N((1\ 1\ \text{-}1)^\top, 0.25I_3)$, and $\beta_i^p \sim N(-3, 1)$. In this random coefficient logit setting without random coefficients, demand inversion with population-level market shares motivates estimation via a nested-fixed point estimation as in Berry et al. (1995) or via the MPEC estimator of Dubé et al. (2012) given in (3.4). Both approaches target identical estimands but the MPEC estimator has computational and numerical advantages (Dubé et al., 2012). Similar to the first simulation, we compare EZ-MPEC to both the infeasible MPEC estimator using the population-level market shares as well as to the feasible MPEC estimator based on estimated market shares that removes product-market combinations with zero-valued market shares. For all

estimators, we evaluate the integral in (3.2) with 200 Monte Carlo draws. Because we do not focus on numerical integration here, we use the same 200 draws in data generation and estimation.

Table 3.2 presents the results for 1,000 simulations. As before, we focus on the median absolute error (MAE) of the mean demand parameter corresponding to the endogenous price variable ($E[\beta_i^p] = -3$). The share of zero-valued market shares in column (4) is the share of the sample dropped from the data when computing the feasible MPEC estimator based on estimated market shares (MPEC: Ad-Hoc) in column (2).

The infeasible MPEC estimator that uses the population-level market shares reported in column (1) of Table 3.2 has an MAE of approximately 0.14. In markets with a small number of consumers between 500-1000, the ad-hoc MPEC has MAEs between approximately 0.263-0.312, highlighting again the relative importance of the sampling uncertainty associated with market shares. In contrast to the setting without random coefficients, the EZ-MPEC estimator does not improve over the ad-hoc MPEC estimator for these small markets. When the number of consumers increases, however, the MAE of the EZ-MPEC estimator decreases at faster rate. We expect that this is due to the fast contraction of the Binomial quantiles as the number of consumers grow, which are used for constructing the feasible set in estimation, relative to the reduction incidental parameter and selection bias that the ad-hoc MPEC estimator suffers from.

Table 3.2: Mean Absolute Error for DGP with Random Coefficients

# Consumers (zero-share)	MPEC: Infeasible	MPEC: Ad-Hoc	EZ-MPEC
	(1)	(2)	(3)
500 (0.091)	0.146	0.312	0.350
750 (0.072)	0.143	0.296	0.298
1000 (0.061)	0.148	0.263	0.250
2000 (0.040)	0.151	0.222	0.205
3000 (0.031)	0.142	0.213	0.189
4000 (0.026)	0.146	0.199	0.178
5000 (0.023)	0.134	0.199	0.171

Notes. Results based on 1,000 Monte Carlo simulations. MPEC: Infeasible and MPEC: Ad-Hoc denote MPEC estimators using the population-level and estimated market shares, respectively. MPEC: Ad-Hoc is computed using only those product-market combinations with positive estimated market shares. Parentheses state the average share of observations with zero-valued estimated market shares.

3.6 Conclusion

This paper proposes a new EZ-MPEC estimator for demand estimation in settings with endogenous prices and estimated market shares. The estimator is constructed by generalizing the constrained optimization formulation of Dubé et al. (2012) for the random coefficient logit model of Berry et al. (1995) using probabilistic bounds on the sampling error of market shares. We show that the estimator is consistent as the number of markets grow large $T \to \infty$ and the number of consumers per market n grows at appropriate rate such that $\log(T)/n \to 0$. Under analogous conditions, we further provide confidence intervals that contain the true demand parameters at pre-specified confidence level.

Two Monte Carlo simulations illustrate the importance of estimation error in market shares and showcase that the incidental parameter and selection problem that conventional estimators suffer from can be substantial. In these settings, application of the proposed EZ-MPEC estimator can lead to meaningful improvements.

3.7 Bibliography

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile Prices in Market Equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Berry, S., O. B. Linton, and A. Pakes (2004). Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies 71*(3), 613–654.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics 25*, 242–262.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica 80*(5), 2231–2267.

Dubé, J.-P., A. Hortaçsu, and J. Joo (2021). Random-coefficients logit demand estimation with zero-valued market shares. *Marketing Science 40*(4), 637–660.

Freyberger, J. (2015). Asymptotic theory for differentiated products demand models with many markets. *Journal of Econometrics 185*(1), 162–181.

Gandhi, A., Z. Lu, and X. Shi (2020). Estimating Demand for Differentiated Products with Zeroes in Market Share Data. *Available at SSRN 3503565*.

Hortaçsu, A., O. R. Natan, H. Parsley, T. Schwieg, and K. R. Williams (2021). Incorporating search

and sales information in demand estimation. Technical report, National Bureau of Economic Research.

Quan, T. W. and K. R. Williams (2018). Product variety, across-market demand heterogeneity, and the value of online retail. *The RAND Journal of Economics 49*(4), 877–913.

Tebaldi, P., A. Torgovitsky, and H. Yang (2019). Nonparametric estimates of demand in the california health insurance exchange. Technical report, National Bureau of Economic Research.

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

3.8 Appendix

3.8.1 Proof of Proposition 51

We begin by establishing the following auxiliary result.

Lemma 56. Fix $n \in \mathbb{N}_{++}$ and $\tau \in (0,1)$. Let $(Z_i)_{i\in\{1,\dots,n\}}$ be a sequence of random variables such that

1. $(Z_i)_{i\in\{1,\dots,n\}}$ is a family of independent random variables,
2. for all i, $\operatorname{supp} Z_i \subset [0,1]$.

Then

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right\| \geqslant \sqrt{\frac{\log\left(\frac{2}{\tau}\right)}{2n}}\right) \leqslant \tau.$$

Proof of Lemma 56. Theorem 2.2.6 in Vershynin (2018) with $M_i = 1/n$, $m_i = 0$ implies that for any $t > 0$,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \mathbb{E}[Z_i]\right) \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}\left(\frac{1}{n}-0\right)^2}\right) = \exp\left(-2t^2 n\right).$$

Then,

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \mathbb{E}[Z_i]\right)\right\| \geqslant t\right) \\
=&\Pr\left(\left\|\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \mathbb{E}[Z_i]\right)\right\| \geqslant t\right\} \bigcup \left\{\frac{1}{n}\sum_{i=1}^{n}\left(-Z_i - \mathbb{E}[-Z_i]\right) \geqslant t\right\}\right) \\
\leqslant& 2\exp\left(-2t^2 n\right)
\end{aligned}$$

where the last inequality applies a union bound. Finally, the right hand side is equal to τ if $t = \sqrt{\frac{\log\left(\frac{2}{\tau}\right)}{2n}}$. □

Proof of Proposition 51. Let $C^j_{n,T}(X, \xi; \theta, \tau)$ denote either

$$\left[\pi_j(X,\xi;\theta) - \sqrt{\frac{\log\left(\frac{2}{\tau}\right)}{2n}}, \pi_j(X,\xi;\theta) + \sqrt{\frac{\log\left(\frac{2}{\tau}\right)}{2n}}\right],$$

or

$$\left[\frac{1}{n}F^{-1}_{\text{Bin}}\left(\frac{\tau}{2}, \pi_j(X,\xi;\theta), n\right), \frac{1}{n}F^{-1}_{\text{Bin}}\left(1-\frac{\tau}{2}, \pi_j(X,\xi;\theta), n\right)\right].$$

Note that by Assumption 8-10, $E[\hat{S}^{(n)}_{jt}] = \pi_j(X_t, \xi_t; \theta_0)$, for all j and t. By Lemma 56 and the definition of quantiles, it follows that in either case

$$\Pr\left(\hat{S}_{jt} \in C^j_{n,T}(X_t, \xi_t; \theta_0, \tau)\right) \geqslant 1 - \tau. \tag{3.7}$$

We then have

$$\begin{aligned}
\Pr\left(\forall j,t:\ \hat{S}_{j,t}\in C^j_{n,T}(X_t,\xi_t;\theta_0,\tau)\right) &= \Pr\left(\bigcap_{t=1}^{T}\bigcap_{j=1}^{J}\left\{\hat{S}_{jt}\in C^j_{n,T}(X_t,\xi_t;\theta_0,\tau)\right\}\right)\\
&\overset{[1]}{=}\prod_{t=1}^{T}\Pr\left(\bigcap_{j=1}^{J}\left\{\hat{S}_{jt}\in C^j_{n,T}(X_t,\xi_t;\theta_0,\tau)\right\}\right)\\
&=\prod_{t=1}^{T}\left[1-\Pr\left(\bigcup_{j=1}^{J}\left\{\hat{S}_{jt}\notin C^j_{n,T}(X_t,\xi_t;\theta_0,\tau)\right\}\right)\right]\\
&\overset{[2]}{\geqslant}\prod_{t=1}^{T}\left[1-\sum_{j=1}^{J}\Pr\left(\hat{S}_{jt}\notin C^j_{n,T}(X_t,\xi_t;\theta_0,\tau)\right)\right]\\
&\overset{[3]}{\geqslant}\left[1-J\tau\right]^T,
\end{aligned}$$

where [1] follows from Assumption 11, [2] follows from the union bound, and [3] follows from inequality (3.7) whenever $J\tau \leqslant 1$. Finally, setting the right hand side equal to $1-\alpha$ and solving for τ, we have $\frac{1-\sqrt[T]{1-\alpha}}{J} \geqslant \tau$. □

3.8.2 Proof of Theorem 53

Proof. The proof proceeds as follows. First, we state an equivalent formulation of the EZ-MPEC estimator using the demand inversion of Berry (1994). Second, we show that consistency of the estimator is implied by consistency in the latent demand shocks. Finally, we show consistency in the latent demand shocks.

Remark 57. The second step of the proof, showing that consistency of the demand estimator is is implied by consistency in the latent demand shocks, relies heavily on the proof of Freyberger (2015), who shows consistency when population-level market shares are replaced by their estimates. We adapt the proof to allow for indeterminacy due to the set-constraints but follow the same arguments elsewhere. Further, when proving consistency in the latent demand shocks, we leverage a lemma that follows from the proof of Freyberger (2015).

3.8.2.1 An Equivalent EZ-MPEC

By Berry et al. (1995), for any realization of the product characteristics and a given value of the demand parameters θ, there exists a bijective map between sample shares and latent demand shocks ξ. Let this map be denoted by $\xi(\cdot)$. For notational convenience, let $\xi_t(s;\theta) \equiv \xi(X_t, s;\theta)$ and similarly $\pi_t(\xi;\theta) \equiv \pi(X_t, s;\theta)$.

Define for any $\alpha \in (0,1)$

$$\mathcal{S}_{n,T}^t(\alpha) = \{s_t \in (0,1)^J : s_t \in C_{n,T}^t(\alpha),\ \|s_t\|_1 < 1\}$$
$$\mathcal{S}_{n,T}(\alpha) = \{\mathbf{s}_T \in (0,1)^{J\times T} : s_t \in \mathcal{S}_{n,T}^t(\alpha), \forall t = 1,\ldots,T\},$$

where

$$C_{n,T}^t(\alpha) \equiv \left[\hat{S}_t^{(n)} - \delta_{n,T}(\alpha),\ \hat{S}_t^{(n)} + \delta_{n,T}(\alpha)\right]$$
$$\delta_{n,T}(\alpha) \equiv \sqrt{\frac{\log\left(\frac{2JT}{\alpha}\right)}{2n}}$$

where $\hat{S}_t^{(n)}$ is the vector of estimated market shares in market t with n consumers.

Let G_T denote the objective function in a sample of size T defined by

$$G_T(\theta, \mathbf{s}_T) = \left(\frac{1}{T}\sum_{t=1} Z_t^\top \xi_t(s_t;\theta)\right)^\top W_T \left(\frac{1}{T}\sum_{t=1} Z_t^\top \xi_t(s_t;\theta)\right).$$

The EZ-MPEC estimator in (3.5) is then equivalent to

$$\begin{aligned} \min_{\theta\in\Theta,\, \mathbf{s}_T\in(0,1)^{J\times T}} \quad & G_T(\theta, \mathbf{s}_T) \\ \text{s.t.} \quad & \mathbf{s}_T \in \mathcal{S}_{n,T}(\alpha), \end{aligned} \tag{3.8}$$

3.8.2.2 Consistency implied by Consistency in Latent Demand Shocks

First, we show that any estimator $\check{\theta}$ is consistent if

$$\|G_T(\check{\theta}, \mathbf{S}_T)\| = \inf_{\theta \in \Theta} \|G_T(\theta, \mathbf{S}_T)\| + o_p(1). \tag{3.9}$$

Fix $\delta > 0$. Note that

$$\begin{aligned}
\Pr\left(\|\check{\theta} - \theta_0\| \geqslant \delta\right) =& \Pr\left(\|\check{\theta} - \theta_0\| \geqslant \delta,\ \|G_T(\check{\theta}, \mathbf{S}_T) - G_T(\theta_0, \mathbf{S}_T)\| \geqslant C(\delta)\right) \\
&+ \Pr\left(\|\check{\theta} - \theta_0\| \geqslant \delta,\ \|G_T(\check{\theta}, \mathbf{S}_T) - G_T(\theta_0, \mathbf{S}_T)\| < C(\delta)\right) \\
\leqslant& \Pr\left(\|G_T(\check{\theta}, \mathbf{S}_T) - G_T(\theta_0, \mathbf{S}_T)\| \geqslant C(\delta)\right) \\
&+ \Pr\left(\inf_{\theta \notin N_{\theta_0}(\delta)} \|G_T(\theta, \mathbf{S}_T) - G_T(\theta_0, \mathbf{S}_T)\| < C(\delta)\right),
\end{aligned}$$

where by Assumption 15

$$\lim_{T \to \infty} \Pr\left(\inf_{\theta \notin N_{\theta_0}(\delta)} \|G_T(\theta, \mathbf{S}_T) - G_T(\theta_0, \mathbf{S}_T)\| < C(\delta)\right) = 0.$$

For the first term, it holds that

$$\begin{aligned}
\|G_T(\check{\theta}, \mathbf{S}_T) - G_T(\theta_0, \mathbf{S}_T)\| &\leqslant \|G_T(\check{\theta}, \mathbf{S}_T)\| + \|G_T(\theta_0, \mathbf{S}_T)\| \\
&\overset{[1]}{=} \|G_T(\theta_0, \mathbf{S}_T)\| + \inf_{\theta \in \Theta} \|G_T(\theta, \mathbf{S}_T)\| + o_p(1) \\
&\leqslant 2\|G_T(\theta_0, \mathbf{S}_T)\| + o_p(1),
\end{aligned}$$

where [1] applies Equation (3.9). Further, by the discussion in Appendix C of Freyberger (2015), assumption 11, 12 and 13, imply that the support of ξ_t is bounded. Using in addition 14, Kol-

mogorov's law of large numbers gives $\|G_T(\theta_0, \mathbf{S}_T)\| = o_p(1)$. Combining, we thus have

$$\Pr\left(\|\check{\theta} - \theta_0\| \geqslant \delta\right) = o_p(1)$$

for any $\check{\theta}$ such that (3.9).

Next, we show that if

$$\sup_{\theta \in \Theta} \left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\theta, \tilde{\mathbf{s}}) \right] - G_T(\theta, \mathbf{S}_T) \right\| = o_p(1) \tag{3.10}$$

then (3.9) holds.

Take any $(\theta_T)_{T=1}^{\infty}$ with $\theta_T \in \Theta$ and note

$$\begin{aligned} \left\| \left\| \inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\theta_T, \tilde{\mathbf{s}}) \right\| - \left\| G_T(\theta_T, \mathbf{S}_T) \right\| \right\| &\leqslant \left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\theta_T, \tilde{\mathbf{s}}) \right] - G_T(\theta_T, \mathbf{S}_T) \right\| \\ &\leqslant \sup_{\theta \in \Theta} \left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\theta, \tilde{\mathbf{s}}) \right] - G_T(\theta, \mathbf{S}_T) \right\| \\ &= o_p(1), \end{aligned} \tag{3.11}$$

where the last equation applies Equation (3.10).

Now define

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \|G_T(\theta, \mathbf{S}_T)\|$$

and

$$\hat{\theta} \in \arg\inf_{\theta \in \Theta,\, \tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\theta, \tilde{\mathbf{s}}).$$

Then

$$
\begin{aligned}
0 \leqslant & \|G_T(\hat{\theta}, \mathbf{S}_T)\| - \inf_{\theta \in \Theta} \|G_T(\theta, \mathbf{S}_T)\| \\
= & \|G_T(\hat{\theta}, \mathbf{S}_T)\| - \|G_T(\tilde{\theta}, \mathbf{S}_T)\| \\
= & \|G_T(\hat{\theta}, \mathbf{S}_T)\| - \left\| \inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\hat{\theta}, \tilde{\mathbf{s}}) \right\| + \left\| \inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\hat{\theta}, \tilde{\mathbf{s}}) \right\| - \|G_T(\tilde{\theta}, \mathbf{S}_T)\| \\
\overset{[1]}{=} & \left\| \inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\hat{\theta}, \tilde{\mathbf{s}}) \right\| - \|G_T(\tilde{\theta}, \mathbf{S}_T)\| + o_p(1) \\
\overset{[2]}{\leqslant} & \left\| \inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\tilde{\theta}, \tilde{\mathbf{s}}) \right\| - \|G_T(\tilde{\theta}, \mathbf{S}_T)\| + o_p(1) \\
\overset{[3]}{=} & o_p(1),
\end{aligned}
$$

where [1] and [3] apply Equation (3.11), and [2] uses the definition of $\hat{\theta}$. Combining, we thus have that (3.9) holds.

By the above, it thus suffices to show that (3.10) holds. Let Z be the $JT \times d_Z$ matrix of instruments. By the Cauchy-Schwarz inequality

$$
\begin{aligned}
\left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} G_T(\theta, \tilde{\mathbf{s}}) \right] - G_T(\theta, \mathbf{S}_T) \right\|^2 &= \frac{1}{T^2} \left\| Z^\top \left(\left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} \xi_T(\tilde{\mathbf{s}}; \theta) \right] - \xi_T(\mathbf{S}_T; \theta) \right) \right\|^2 \\
&\leqslant \frac{1}{T} \|Z^\top Z\| \times \frac{1}{T} \left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} \xi_T(\tilde{\mathbf{s}}; \theta) \right] - \xi_T(\mathbf{S}_T; \theta) \right\|^2.
\end{aligned}
$$

Since $\frac{1}{T}\|Z^\top Z\| = O_p(1)$ by Assumption 14, it suffices to prove that the second term is $o_p(1)$. For this purpose, note further that

$$
\begin{aligned}
\sup_{\theta \in \Theta} \frac{1}{T} \left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}} \xi_T(\tilde{\mathbf{s}}; \theta) \right] - \xi_T(\mathbf{S}_T; \theta) \right\|^2 &= \sup_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \left\| \left[\inf_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}^t} \xi_t(\tilde{\mathbf{s}}_t; \theta) \right] - \xi_T(\mathbf{S}_T; \theta) \right\|^2 \\
&\leqslant \sup_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}^t} \left\| \xi_t(\tilde{\mathbf{s}}_t; \theta) - \xi_T(\mathbf{S}_T; \theta) \right\|^2 \qquad (3.12) \\
&\leqslant \sup_{\theta \in \Theta} \max_{t \in \{1, \ldots, T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}_{n,T}^t} \left\| \xi_t(\tilde{\mathbf{s}}_t; \theta) - \xi_T(\mathbf{S}_T; \theta) \right\|^2,
\end{aligned}
$$

so that we may consider the final expression.

3.8.2.3 Consistency of Latent Demand Shocks

For $\alpha \in (0,1)$, let $\mathcal{H}_{n,T}(\alpha)$ denote the the event that all Hoeffding bounds on the sampling error hold in the sample – that is,

$$\mathcal{H}_{n,T}(\alpha) \equiv \left\{ \pi(X_t, \xi_t; \theta_0) \in C^t_{n,T}(\alpha),\ \forall t = 1, \ldots, T \right\}.$$

Now notice that by the definition of $\pi(\cdot)$ and $\xi(\cdot)$, it holds for any $C(\epsilon) > 0$ that

$$\begin{aligned}
&\Pr\left(\sup_{\theta \in \Theta} \max_{t \in \{1,\ldots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \pi_t(\xi_t(\tilde{\mathbf{s}}_t; \theta); \theta) - \pi_t(\xi_t(\mathbf{S}_t; \theta); \theta) \right\| \geqslant C(\epsilon) \right) \\
&\overset{[1]}{=} \Pr\left(\max_{t \in \{1,\ldots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \tilde{\mathbf{s}}_t - \mathbf{S}_t \right\| \geqslant C(\epsilon) \right) \\
&\overset{[2]}{=} \Pr\left(\max_{t \in \{1,\ldots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \tilde{\mathbf{s}}_t - \mathbf{S}_t \right\|^2 \geqslant C(\epsilon) \middle| \mathcal{H}_{n,T}(\alpha) \right) \Pr\left(\mathcal{H}_{n,T}(\alpha) \right) \\
&\qquad + \Pr\left(\max_{t \in \{1,\ldots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \tilde{\mathbf{s}}_t - \mathbf{S}_t \right\|^2 \geqslant C(\epsilon) \middle| (\mathcal{H}_{n,T}(\alpha))^c \right) \Pr\left((\mathcal{H}_{n,T}(\alpha))^c \right) \\
&\overset{[3]}{\leqslant} \Pr\left(\max_{t \in \{1,\ldots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \tilde{\mathbf{s}}_t - \mathbf{S}_t \right\|^2 \geqslant C(\epsilon) \middle| \mathcal{H}_{n,T}(\alpha) \right) + \Pr\left((\mathcal{H}_{n,T}(\alpha))^c \right) \\
&\overset{[4]}{\leqslant} \Pr\left(\max_{t \in \{1,\ldots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \tilde{\mathbf{s}}_t - \mathbf{S}_t \right\|^2 \geqslant C(\epsilon) \middle| \mathcal{H}_{n,T}(\alpha) \right) + \alpha \\
&\overset{[5]}{\leqslant} \Pr\left(2\delta_{n,T}(\alpha) \geqslant C(\epsilon) | \mathcal{H}_{n,T}(\alpha) \right) + \alpha \\
&\overset{[6]}{\leqslant} \mathbb{1}\{2\delta_{n,T}(\alpha) \geqslant C(\epsilon)\} + \alpha
\end{aligned} \tag{3.13}$$

where [1] uses that the bounds $\mathcal{S}^t_{n,T}$ based on Hoeffding's inequality do not depend on θ, [2] follows from the law of total probability, [3] uses that probabilities are bounded by one, [4] follows from Proposition 51 which implies $\Pr\left(\mathcal{H}_{n,T}\right) \geqslant 1 - \alpha$, [5] follows from the definition of $\mathcal{S}^t_{n,T}$, and [6] follows since $\delta_{n,T}(\alpha)$ is non-random. Choosing $\alpha = \alpha_{n,T} = o_p(1)$, it then follows from the

definition of $\delta_{n,T}(\alpha)$ that the term converges to zero whenever $\log(T) = o_p(n)$.

This implies that (3.12) is $o_p(1)$. To see this, suppose by way of contradiction that for some $\delta > 0$

$$\sup_{\theta \in \Theta} \max_{t \in \{1,\dots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \xi_t(\tilde{\mathbf{s}}_t; \theta) - \xi_T(\mathbf{S}_T; \theta) \right\|^2 > \delta.$$

Then by Lemma 58, which we state below, there exists $C(\delta) > 0$ such that

$$\sup_{\theta \in \Theta} \max_{t \in \{1,\dots,T\}} \sup_{\tilde{\mathbf{s}} \in \mathcal{S}^t_{n,T}} \left\| \pi_t(\xi_t(\tilde{\mathbf{s}}_t; \theta); \theta) - \pi_t(\xi_t(\mathbf{S}_t; \theta); \theta) \right\| > C(\delta),$$

which contradicts that (3.13) is $o_p(1)$ whenever $\log(T) = o_p(n)$. Hence, we have the desired result. □

3.8.2.4 A Useful Lemma

Lemma 58. Under the assumptions of Theorem 53, it holds that $\forall \delta > 0$, $\exists C(\delta) > 0$ such that for all $t = 1, \dots, T$,

$$\Pr\left(\inf_{\theta \in \Theta} \inf_{\tilde{\xi} \notin \mathcal{N}_{\xi_t}(\delta)} \left\| \pi_t(\tilde{\xi}; \theta) - \pi_t(\xi_t; \theta) \right\| > C(\delta) \right) = 1.$$

Remark 59. The lemma follows from the proof of Theorem 1 in Freyberger (2015) with only minor adaptations. We include it here for completeness.

Proof. Take $\tilde{\xi} : \|\tilde{\xi} - \xi_t\| \geqslant J\delta$. Without loss of generality, take $\|\tilde{\xi}_1 - \xi_{1t}\| \geqslant \|\tilde{\xi}_j - \xi_{jt}\|, \forall j = 1, \dots, J$. We further take $\tilde{\xi}_1 - \xi_{1t} \geqslant \delta$, but the arguments for $\tilde{\xi}_1 - \xi_{1t} \leqslant -\delta$ are analogous.

Let $\tilde{\xi} + \delta$ denote element-wise addition. For all $\delta > 0$, it holds that

$$\exp(\delta)\pi_{1t}(\xi_t; \theta) > \pi_{1t}(\xi_t + \delta; \theta) > \pi_{1t}(\xi_t; \theta), \quad \text{and} \quad \pi_{1t}(\tilde{\xi}; \theta) \geqslant \pi_{1t}(\xi_t + \delta; \theta). \tag{3.14}$$

Further, by assumptions 8, 12, 13, and 11, there exists $\gamma > 0$ and a compact set $\mathcal{V} \subset \mathbb{R}^{d_X}$ such that $\int_{\mathcal{V}} dF(\beta;\theta) > p > 0$ and with probability 1,

$$\gamma < \Pr(Y_i = j \| p_t, X_t, \xi_t, \beta) < 1 - \gamma, \tag{3.15}$$

for all $j = 0, 1, \ldots, J$, and $\beta \in \mathcal{V}$, where

$$\Pr(Y_i = j \| p, X, \xi; \beta) = \frac{\exp\left(V(p_j, X_j; \beta) + \xi_j\right)}{1 + \sum_{l=1}^{J} \exp\left(V(p_l, X_l; \beta) + \xi_l\right)}.$$

For notational convenience, let $v_{jt}(\xi;\beta) = \Pr(Y_i = j \| p_t, X_t, \xi; \beta)$

Next, define $\delta_0 = \min\{\delta, -\frac{1}{4}\log(1-\gamma)\}$. We have

$$\begin{aligned}
&\int v_{1t}(\tilde{\xi};\beta)dF(\beta;\theta) - \int v_{1t}(\xi_t;\beta)dF(\beta;\theta) \\
\geqslant &\int_{\mathcal{V}} v_{1t}(\tilde{\xi};\beta)dF(\beta;\theta) - \int_{\mathcal{V}} v_{1t}(\xi_t;\beta)dF(\beta;\theta) \\
\geqslant &\int_{\mathcal{V}} v_{1t}(\xi_t + \delta_0;\beta)dF(\beta;\theta) - \int_{\mathcal{V}} v_{1t}(\xi_t;\beta)dF(\beta;\theta).
\end{aligned}$$

By the mean value theorem, it then holds that for some $\tilde{\delta} \in (0, \delta_0)$

$$\begin{aligned}
&\int_{\mathcal{V}} v_{1t}(\xi_t + \delta_0;\beta)dF(\beta;\theta) - \int_{\mathcal{V}} v_{1t}(\xi_t;\beta)dF(\beta;\theta) \\
= &\left(\int_{\mathcal{V}} v_{1t}(\xi_t + \tilde{\delta};\beta)dF(\beta;\theta) - \int_{\mathcal{V}} v_{1t}(\xi_t + \tilde{\delta};\beta)\left(\sum_{k=1}^{J} v_{kt}(\xi_t + \tilde{\delta};\beta)\right)dF(\beta;\theta)\right)\delta_0, \\
\overset{[1]}{\geqslant} &\left(\int_{\mathcal{V}} v_{1t}(\xi_t;\beta)dF(\beta;\theta) - \exp(2\tilde{\delta})\int_{\mathcal{V}} v_{1t}(\xi_t;\beta)\left(\sum_{k=1}^{J} v_{kt}(\xi_t;\beta)\right)dF(\beta;\theta)\right)\delta_0 \\
\overset{[2]}{\geqslant} &\int_{\mathcal{V}} v_{1t}(\xi_t;\beta)dF(\beta;\theta)\left(1 - \exp(2\tilde{\delta})(1-\gamma)\right)\delta_0 \\
\overset{[3]}{\geqslant} &\int_{\mathcal{V}} dF(\beta;\theta)\gamma\left(1 - \exp(2\tilde{\delta})(1-\gamma)\right)\delta_0 \\
\overset{[4]}{\geqslant} &p\gamma\left(1 - \exp(2\tilde{\delta})(1-\gamma)\right)\delta_0,
\end{aligned}$$

where [1] follows from (3.14), [2] and [3] follows from (3.15), and [3] uses that $\int_{\mathcal{V}} dF(\beta;\theta) > p > 0$. The last term is greater than zero and only depends on δ. Hence, we can take $C(\delta) = p\gamma\left(1-\exp(2\tilde{\delta})(1-\gamma)\right)\delta_0$ which completes the proof. □

3.8.3 Proof of Theorem 55

Proof. In addition to the test statistic $\hat{G}_T(\alpha)$ defined in (3.6) using the sampling bounds used in estimation, we define two additional test statistics. The first is based on *multinomial* quantile bounds $\tilde{C}_{n,T}(X,\xi;\theta,\alpha)$ with the property that

$$\Pr\left(\tilde{\mathcal{H}}_{n,T}(\alpha)\right) = \alpha, \tag{3.16}$$

where $\tilde{\mathcal{H}}_{n,T}(\alpha)$ denotes the the event that all multinhatomial bounds on the sampling error hold in the sample – that is,

$$\tilde{\mathcal{H}}_{n,T}(\alpha) \equiv \left\{\hat{S}_t^{(n)} \in \tilde{C}_{n,T}(X,\xi_t;\theta_0,\alpha),\ \forall t = 1,\ldots,T\right\}.$$

In particular, define

$$\begin{aligned}\tilde{G}_T(\alpha) \equiv \min_{\{\tilde{\xi}_t\}_{t=1}^T} \quad & T\left(\frac{1}{T}\sum_{t=1}^T Z_t^\top \tilde{\xi}_t\right)^\top \left(\frac{1}{T}\sum_{t=1}^T \hat{\xi}_t^\top Z_t Z_t^\top \hat{\xi}_t\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T Z_t^\top \tilde{\xi}_t\right) \\ \text{s.t.} \quad & \hat{S}_t^{(n)} \in \tilde{C}_{n,T}(X,\tilde{\xi}_t;\theta_0,\alpha),\ \forall t \in \{1,\ldots,T\}.\end{aligned} \tag{3.17}$$

Note that by construction,

$$\tilde{C}_{n,T}(X,\xi_t;\theta_0,\alpha) \subset C_{n,T}(X,\xi_t;\theta_0,\alpha),\ \forall t \in \{1,\ldots,T\},$$

where the RHS are the bounds specified in Proposition 51 that we consider in estimation, so that

$$\Pr\left(\hat{G}_T(\alpha) \leqslant \tilde{G}_T(\alpha)\right) = 1. \tag{3.18}$$

Further, define the infeasible GMM test statistic

$$G_T \equiv T\left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right)^\top \left(\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}_t^\top Z_t Z_t^\top \hat{\xi}_t\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t\right),$$

where ξ_t is the set of true latent demand shocks. Note then that under H_0,

$$\Pr\left(\tilde{G}_T(\alpha) \leqslant G_T(\alpha)\middle|\tilde{\mathcal{H}}_{n,T}(\alpha)\right) = 1. \tag{3.19}$$

Then, let $c_K^{1-\tau}$ denote the $1-\tau$th quantile of a χ^2 distribution with K degrees of freedom

$$\begin{aligned}
E\left[\mathbb{1}\{\hat{G}_T(\alpha) > c_K^{1-\tau}\}\right] &\overset{[1]}{\leqslant} E\left[\mathbb{1}\{\tilde{G}_T(\alpha) > c_K^{1-\tau}\}\right] \\
&= E\left[\mathbb{1}\{\tilde{G}_T(\alpha) > c_K^{1-\tau}\}\middle|\tilde{\mathcal{H}}_{n,T}(\alpha)\right]\Pr\left(\tilde{\mathcal{H}}_{n,T}(\alpha)\right) \\
&\quad + E\left[\mathbb{1}\{\tilde{G}_T(\alpha) > c_K^{1-\tau}\}\middle|(\tilde{\mathcal{H}}_{n,T}(\alpha))^c\right]\Pr\left((\tilde{\mathcal{H}}_{n,T}(\alpha))^c\right) \\
&\overset{[2]}{\leqslant} E\left[\mathbb{1}\{\tilde{G}_T(\alpha) > c_K^{1-\tau}\}\middle|\tilde{\mathcal{H}}_{n,T}(\alpha)\right](1-\alpha) + \alpha \\
&\overset{[3]}{\leqslant} E\left[\mathbb{1}\{G_T > c_K^{1-\tau}\}\middle|\tilde{\mathcal{H}}_{n,T}(\alpha)\right](1-\alpha) + \alpha \\
&\overset{[4]}{=} E\left[\mathbb{1}\{G_T > c_K^{1-\tau}\}\right](1-\alpha) + \alpha,
\end{aligned} \tag{3.20}$$

where [1] follows from (3.18), [2] follows from (3.16), and [3] follows from (3.19). To show step [4], we can show that

$$\{(Z_t, \xi_t)\}_{t=1}^T \perp\!\!\!\perp \tilde{\mathcal{H}}_{n,T}(\alpha),$$

since G_T is a function of only $\{(Z_t, \xi_t)\}_{t=1}^T$. To do so, let A_T be any set on $\mathbb{R}^{T\times J\times(K+1)}$ and

consider

$$
\begin{aligned}
E\left[\mathbb{1}_{A_T}\left(\{(Z_t,\xi_t\}_{t=1}^T\right)\mathbb{1}\{\tilde{\mathcal{H}}_{n,T}(\alpha)\}\right] &= E\left[\mathbb{1}_{A_T}\left(\{(Z_t,\xi_t\}_{t=1}^T\right)E\left[\mathbb{1}\{\tilde{\mathcal{H}}_{n,T}(\alpha)\}\middle|\{(Z_t,\xi_t,S_t\}_{t=1}^T\right]\right] \\
&= E\left[\mathbb{1}_{A_T}\left(\{(Z_t,\xi_t\}_{t=1}^T\right)E\left[\mathbb{1}\{\tilde{\mathcal{H}}_{n,T}(\alpha)\}\middle|\{S_t\}_{t=1}^T\right]\right] \\
&= E\left[\mathbb{1}_{A_T}\left(\{(Z_t,\xi_t\}_{t=1}^T\right)(1-\alpha)\right] \\
&= E\left[\mathbb{1}_{A_T}\left(\{(Z_t,\xi_t\}_{t=1}^T\right)\right]E\left[\mathbb{1}\{\tilde{\mathcal{H}}_{n,T}(\alpha)\}\right].
\end{aligned}
$$

Freyberger (2015) shows that under assumptions 8, 12, and 13, the latent demand shocks ξ_t are bounded. Then, by assumptions 9, 11, and 14, it holds by the central limit theorem that $\sqrt{T}\frac{1}{T}\sum_{t=1}^T Z_t^\top \xi_t \xrightarrow{d} N(0,\Sigma)$, where ξ_t are the true latent demand shocks. Hence by Theorem 53 and the continuous mapping theorem

$$
G_T \equiv T\left(\frac{1}{T}\sum_{t=1}^T Z_t^\top \xi_t\right)^\top \left(\frac{1}{T}\sum_{t=1}^T \hat{\xi}_t^\top Z_t Z_t^\top \hat{\xi}_t\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T Z_t^\top \xi_t\right) \xrightarrow{d} \chi^2(K), \tag{3.21}
$$

where $\chi^2(K)$ is a χ^2 distribution with K degrees of freedom. Combining (3.20) and (3.21) and using that $\alpha_{n,T} = o_p(1)$, we have under H_0,

$$
\limsup_{t\to\infty} E\left[\mathbb{1}\{\hat{G}_T(\alpha) > c_K^{1-\tau}\}\right] \leqslant \limsup_{T\to\infty} E\left[\mathbb{1}\{G_T > c_K^{1-\tau}\}\right](1-\alpha_{n,T}) + \alpha_{n,T} = \tau.
$$

□

3.8.4 Nonparametric Demand Estimation

While the zero-share problem is particularly salient for the random coefficients logit model, the underlying problem of errors in estimated choice probabilities exists in many frameworks for demand estimation. To illustrate this, we extend the nonparametric demand estimation framework proposed by Tebaldi et al. (2019) to a finite number of consumers. For computational tractability, this model requires linearity of any additional constraints. We therefore focus on using Hoeffding's

inequality.

In the simpler framework with exogenous prices, the linear program is[9]

$$\max_{\phi} / \min_{\phi} c'\phi \qquad \begin{cases} -\phi_k & \leqslant 0, \\ \sum_k \phi_k & = 1, \\ \sum_{k \in S(j,m)} \phi_k & = \hat{s}_{j,m} \quad \forall j, m. \end{cases} \tag{3.22}$$

where $S(j,m)$ is the set of indices of elements of the Minimum Relevant Partition whose union is the market share for product j in market m. $S(j,m)$ is a deterministic function of prices and known to the researcher. The only random variable here is $\hat{s}_{j,m}$, the estimated market share. We can apply Hoeffding's inequality to obtain simultaneous confidence intervals for all products in all markets. This leads to

$$\max_{\phi} / \min_{\phi} c'\phi \qquad \begin{cases} -\phi_k & \leqslant 0, \\ \sum_k \phi_k & = 1, \\ \left\| \sum_{k \in S(j,m)} \phi_k - \hat{s}_{j,m} \right\| & \leqslant B_{(n_t)_t, \alpha} \quad \forall j, m. \end{cases} \tag{3.23}$$

Note that the absolute value in the concentration constraints can be written as linear constraints. Hence our modification (3.23) of (3.22) maintains computational tractability since efficient solvers for linear programs are available.

One interesting feature of (3.23) and its generalization to endogenous prices proposed by Tebaldi et al. (2019) is that there often does not exist a feasible solution in applications. This can be interpreted as a rejection of the econometric model by (3.22), notably the quasi-linearity of prices or the time-market-homogeneity of latent utility draws. However, the reason why the model appears rejected could be the finiteness of consumers. While we do not have access to the data

[9]The goal is to bound the counterfactual from both sides, hence the $\max \min$ notation.

used in Tebaldi et al. (2019), we can replicate this phenomenon in simulations.

3.8.5 Implementation details

This section outlines implementation details of the EZ-MPEC estimator using Binomial quantiles and supplies expressions for the Jacobian and Hessians used in estimation. All programs are implemented in Matlab using Knitro, with code readily available upon request.

3.8.5.1 Setup

Note that using the bounds derived in Proposition 51, we can write the corresponding EZ-MPEC estimator as

$$\begin{aligned}
\min_{(\theta,\xi,\eta)} \quad & \eta^\top W \eta \\
\text{s.t.} \quad & \frac{1}{n} F_{\text{Bin}}^{-1}\left(\frac{1-\sqrt[T]{1-\alpha}}{J}, P_{jt}(\theta), n\right) - \hat{S}_{j,t} \leqslant 0, \ \forall j,t, \\
& \hat{S}_{j,t} - \frac{1}{n} F_{\text{Bin}}^{-1}\left(1 - \frac{1-\sqrt[T]{1-\alpha}}{J}, P_{jt}(\theta), n\right) \leqslant 0, \ \forall j,t, \\
& \eta = \frac{1}{T}\sum_{t=1}^{T} Z_t^\top \xi_t
\end{aligned}$$

where $P_{jt}(\theta) \equiv \Pr(Y_i = j \| p_t, X_t, \xi_t, \theta)$ as given in equation (3.2). The task is now to obtain the Jacobian of the objective function and the constraints, as well as the Hessian of the corresponding Lagrangian. Fortunately, the program is similar to Dubé et al. (2012), which derive first and second order derivatives of the objective, the moment equality, as well as the CCPs $P_{jt}(\theta)$. We thus focus on the derivatives of the Binomial quantile function with respect to the CCPs $P_{jt}(\theta)$.

Since the Binomial quantile function is not continuously differentiable, we approximate the Binomial distribution function with a function that has a continuously differentiable inverse. We

consider

$$\hat{F}(r, P_{jt}(\theta), n) \equiv \sum_{i=0}^{n} \frac{1}{1 + \exp(-2m(r-i))} \binom{n}{i} P_{jt}(\theta)^i \left(1 - P_{jt}(\theta)\right)^{n-i}, \tag{3.24}$$

where $m \geqslant 0$ is a hyperparameter.[10]

3.8.5.2 Inverse function rules

Although the approximation (3.24) has an inverse in its first argument that is continuously differentiable in its second argument, it does generally have a closed form expression. We thus leverage simple results akin to the inverse function theorem.

In particular, we have

$$\begin{aligned} & \hat{F}\left(\hat{F}^{-1}(\alpha, p, n), p, n\right) = \alpha \\ \Rightarrow \quad & \frac{\partial}{\partial} \hat{F}\left(\hat{F}^{-1}(\alpha, p, n), p, n\right) = 0, \end{aligned}$$

and hence by application of the chain rule and the inverse function theorem

$$\begin{aligned} 0 &= \frac{\partial}{\partial} \hat{F}\left(\hat{F}^{-1}(\alpha, p, n), p, n\right) \\ &= \frac{\partial \hat{F}(r, p, n)}{\partial r}\Big\|_{r=\hat{F}^{-1}(\alpha,p,n)} \frac{\partial \hat{F}^{-1}(\alpha, p, n)}{\partial p} + \frac{\partial \hat{F}(\hat{F}^{-1}(\alpha, p, n), p, n)}{\partial p} \\ \Rightarrow \quad \frac{\partial \hat{F}^{-1}(\alpha, p, n)}{\partial p} &= \frac{-\frac{\partial \hat{F}(\hat{F}^{-1}(\alpha,p,n),p,n)}{\partial p}}{\frac{\partial \hat{F}(r,p,n)}{\partial r}\Big\|_{r=\hat{F}^{-1}(\alpha,p,n)}}. \end{aligned} \tag{3.25}$$

For the second order derivative, we simplify notation and let subscripts denote partial derivatives with respect to the indexed argument.

[10] To account for the approximation, we can further shift the bounds by a constant. This would not affect the derivatives.

Then

$$\begin{aligned}\frac{\partial^2 \hat{F}^{-1}(\alpha,p,n)}{\partial^2 p} &= \hat{F}_{22}^{-1}(\alpha,p,n) = \frac{\partial}{\partial p}\left(\hat{F}_2^{-1}(\alpha,p,n)\right) = \frac{\partial}{\partial p}\left(\frac{-\hat{F}_2(\hat{F}^{-1}(\alpha,p,n),p,n)}{\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)}\right) \\ &= \frac{\hat{F}_2(\hat{F}^{-1}(\alpha,p,n),p,n)}{\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)^2}\frac{\partial}{\partial p}\left(\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)\right) \\ &\quad - \frac{1}{\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)}\frac{\partial}{\partial p}\left(\hat{F}_2(\hat{F}^{-1}(\alpha,p,n),p,n)\right) \\ &= \frac{1}{\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)}\Bigg(\hat{F}_2^{-1}(\alpha,p,n)\frac{\partial}{\partial p}\left(\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)\right) \\ &\quad - \frac{\partial}{\partial p}\left(\hat{F}_2(\hat{F}^{-1}(\alpha,p,n),p,n)\right)\Bigg),\end{aligned} \tag{3.26}$$

where

$$\begin{aligned}\frac{\partial}{\partial p}\left(\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)\right) &= \hat{F}_{11}(\hat{F}^{-1}(\alpha,p,n),p,n)\hat{F}_2^{-1}(\alpha,p,n) + \hat{F}_{12}(\hat{F}^{-1}(\alpha,p,n),p,n), \\ \frac{\partial}{\partial p}\left(\hat{F}_2(\hat{F}^{-1}(\alpha,p,n),p,n)\right) &= \hat{F}_{21}(\hat{F}^{-1}(\alpha,p,n),p,n)\hat{F}_2^{-1}(\alpha,p,n) + \hat{F}_{22}(\hat{F}^{-1}(\alpha,p,n),p,n).\end{aligned}$$

Note that by Schwarz's theorem, $\hat{F}_{12}(x,p,n) = \hat{F}_{21}(x,p,n)$. Hence, Equation (3.26) simplifies to

$$\begin{aligned}\frac{\partial^2 \hat{F}^{-1}(\alpha,p,n)}{\partial^2 p} = \frac{1}{\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)}\Bigg(&\hat{F}_{11}(\hat{F}^{-1}(\alpha,p,n),p,n)\left(\hat{F}_2^{-1}(\alpha,p,n)\right)^2 \\ &- \hat{F}_{22}(\hat{F}^{-1}(\alpha,p,n),p,n)\Bigg).\end{aligned} \tag{3.27}$$

3.8.5.3 Jacobian

Using Equation (3.25) and the chain rule, we have

$$\frac{\partial \hat{F}^{-1}(\alpha,P_{j,t}(\theta),n)}{\partial \theta} = \left(\frac{-\frac{\partial \hat{F}(\hat{F}^{-1}(\alpha,P_{j,t}(\theta),n),p,n)}{\partial p}\Big\|_{p=P_{j,t}(\theta)}}{\frac{\partial \hat{F}(r,p,n)}{\partial r}\Big\|_{r=\hat{F}^{-1}(\alpha,P_{j,t}(\theta),n)}}\right)\frac{\partial P_{j,t}(\theta)}{\partial \theta}, \tag{3.28}$$

where $\frac{\partial P_{j,t}(\theta)}{\partial \theta}$ is the same as in Dubé et al. (2012), and

$$\frac{\partial \hat{F}(r,p,n)}{\partial p} = \sum_{i=0}^{n} \frac{\frac{i}{p} - \frac{n-i}{1-p}}{1+\exp(-2m(r-i))} \binom{n}{i} p^i \left(1-p\right)^{n-i},$$

and

$$\frac{\partial \hat{F}(r,p,n)}{\partial r} = 2m \sum_{i=0}^{n} \frac{\exp\left(-2m(r-i)\right)}{\left(1+\exp(-2m(r-i))\right)^2} \binom{n}{i} p^i (1-p)^{n-i}.$$

3.8.5.4 Hessian

Using the chain rule, we have

$$\begin{aligned}\frac{\partial^2 \hat{F}^{-1}(\alpha, P_{j,t}(\theta), n)}{\partial\theta\partial\theta^\top} &= \frac{\partial}{\partial\theta^\top}\left(\frac{\partial \hat{F}^{-1}(\alpha, P_{j,t}(\theta), n)}{\partial\theta}\right) \\ &= \frac{\partial^2 \hat{F}^{-1}(\alpha,p,n)}{\partial p^2}\bigg\|_{p=P_{j,t}(\theta)} \frac{\partial P_{j,t}(\theta)}{\partial\theta}\frac{\partial P_{j,t}(\theta)}{\partial\theta^\top} + \frac{\partial \hat{F}^{-1}(\alpha,p,n)}{\partial p}\bigg\|_{p=P_{j,t}(\theta)} \frac{\partial P_{j,t}(\theta)}{\partial\theta\partial\theta^\top},\end{aligned}$$

where $\frac{\partial P_{j,t}(\theta)}{\partial\theta}$ and $\frac{\partial P_{j,t}(\theta)}{\partial\theta\partial\theta^\top}$ are the same as in Dubé et al. (2012), and $\frac{\partial \hat{F}^{-1}(\alpha,p,n)}{\partial p}\Big\|_{p=P_{j,t}(\theta)}$ is given in equation (3.28). Finally, using Equation (3.27), we have

$$\begin{aligned}\frac{\partial^2 \hat{F}^{-1}(\alpha,p,n)}{\partial p^2} = \frac{1}{\hat{F}_1(\hat{F}^{-1}(\alpha,p,n),p,n)}\Bigg(&\hat{F}_{11}(\hat{F}^{-1}(\alpha,p,n),p,n)\left(\hat{F}_2^{-1}(\alpha,p,n)\right)^2 \\ &- \hat{F}_{22}(\hat{F}^{-1}(\alpha,p,n),p,n)\Bigg),\end{aligned}$$

where

$$\hat{F}_{22}(r,p,n) = \sum_{i=0}^{n} \frac{\frac{i(i-1)}{p^2} - 2\frac{i(n-i)}{\Pr(1-p)} + \frac{(n-i)(n-i-1)}{(1-p)^2}}{1+\exp(-2m(r-i))} \binom{n}{i} p^i \left(1-p\right)^{n-i},$$

and

$$\hat{F}_{11}(r,p,n) = -2m \sum_{i=0}^{n} \frac{\exp\left(-2m(r-i)\right)}{\left(1+\exp(-2m(r-i))\right)^2} \left(1 - 2\frac{\exp\left(-2m(r-i)\right)}{1+\exp(-2m(r-i))}\right) \binom{n}{i} p^i \left(1-p\right)^{n-i},$$

and $\hat{F}_1(r,p,n)$ and $\hat{F}_2^{-1}(\alpha,p,n)$ as before.