Estimating Demand for Differentiated Products with Zeroes in Market Share Data*

Amit Gandhi
UPenn
Microsoft

Zhentong Lu
Bank of Canada

Xiaoxia Shi †
UW-Madison

June 1, 2019

Abstract

In this paper we introduce a new approach to estimating differentiated product demand systems that allows for products with zero sales in the data. Zeroes in demand are a common problem in product differentiated markets, but fall outside the scope of existing demand estimation techniques. Our solution to the zeroes problem is based on constructing bounds for the conditional expectation of the inverse demand. These bounds can be translated into moment inequalities that are shown to yield consistent and asymptotically normal point estimator for demand parameters under natural conditions for differentiated product markets. In Monte Carlo simulations, we demonstrate that the new approach works well even when the fraction of zeroes is as high as 95%. We apply our estimator to supermarket scanner data and find that correcting the bias caused by zeroes has important empirical implications, e.g., price elasticities become on the order of twice as large when zeroes are properly controlled.

Keywords: Demand Estimation, Differentiated Products, Measurement Error, Moment Inequality, Zero

JEL: C01, C12, L10, L81.

1 Introduction

In this paper we introduce a new approach to differentiated product demand estimation that allows for zeroes in empirical market share data. Such zeroes are a highly prevalent feature of demand in

*Previous version of this paper was circulated under the title "Estimating Demand for Differentiated Products with Error in Market Shares."

†We are thankful to Steven Berry, Jean-Pierre Dubé, Philip Haile, Bruce Hansen, Ulrich Müller, Aviv Nevo, Jack Porter, and Chris Taber for insightful discussions and suggestions; We would also like to thank the participants at the MIT Econometrics of Demand Conference, Chicago-Booth Marketing Lunch, the Northwestern Conference on "Junior Festival on New Developments in Microeconometrics," the Cowles Foundation Conference on "Structural Empirical Microeconomic Models," 3rd Cornell - Penn State Econometrics & Industrial Organization Workshop, as well as seminar participants at Wisconsin-Madison, Wisconsin-Milwaukee, Cornell, Indiana, Princeton, NYU, Penn and the Federal Trade Commission for their many helpful comments and questions.

a variety of empirical settings, ranging from workhorse scanner retail data, to data as diverse as homicide rates and international trade flows (we discuss these examples in further depth below). Zeroes naturally arise in "big data" applications which allow for increasingly granular views of consumers, products, and markets (see for example Quan and Williams (2015), Nurski and Verboven (2016)). Unfortunately, the standard estimation procedures following the seminal Berry, Levinsohn, and Pakes (1995) (BLP for short) cannot be used in the presence of zero empirical shares - they are simply not well defined when zeroes are present. Furthermore, ad hoc fixes to market zeroes that are sometimes used in practice, such as dropping zeroes from the data or replacing them with small positive numbers, are subject to biases which can be quite large (discussed further below). This has left empirical work on demand for differentiated products without satisfying solutions to the zero shares problem, and often force researchers to aggregate their rich data on naturally defined products to crude artificial products which limits the type of questions that can be answered. This is the key problem that our paper aims to solve.

In this paper we provide an approach to estimating differentiated product demand models that provides consistency (and asymptotic normality) for demand parameters despite a possibly large presence of zero market shares in the data. We first isolate the econometric problem caused by zeroes in the data. The problem we show is driven by the wedge between *choice probabilities*, which are the theoretical outcome variables predicted by the demand model, and *market shares*, which are the empirical revealed preference data used to estimate choice probabilities. Although choice probabilities are strictly positive in the underlying model, market shares are often zero if *choice probabilities are small*. The root of the zeroes problem is that substituting market shares (or some other consistent estimate) for choice probabilities in the moment conditions that identify the model, which is the basis for the traditional estimators, will generally lead to asymptotic bias. While this bias is assumed away in the traditional approach, it cannot be avoided whenever zeroes are prevalent in the data.

Our solution to this problem is to construct a set of moment *inequalities* for the model, which are by design robust to the sampling error in market shares - our moment inequalities will hold at the true value of the parameters regardless of the magnitude of the error in market shares as a measurement for choice probabilities. Despite taking an inequality form, we use these moment inequalities to form a GMM-type point estimator based on minimizing the deviations from the inequalities. We show this estimator is consistent so long as there is a positive mass of observations whose latent choice probabilities are bounded sufficiently away from zero, e.g., products for whom market shares are not likely to be zero. This is natural in many applications (as illustrated in Section 2), and strictly generalizes the restrictions on choice probabilities for consistency under the traditional approach. Asymptotic normality then follows by similar arguments as those for censored regression models by Kahn and Tamer (2009).

Computationally, our estimator closely resembles the traditional approach with only a slight adjustment in how the empirical moments are constructed. In particular it is no more burdensome than the usual estimation procedures for BLP and can be implemented using either the standard

nested fixed point method of the original BLP, or the MPEC method as advocated more recently by Dubé, Fox, and Su (2012).

We investigate the finite sample performance of the approach in a variety of mixed logit examples. We find that our estimator works well even when the the fraction of zeros is as high as 95%, while the standard procedure with the observations with zeroes deleted yields severely biased estimators even with mild or moderate fractions of zeroes.

We apply our bounds approach to widely used scanner data from the Dominicks Finer Foods (DFF) retail chain. In particular, we estimate demand for the tuna category as previously studied by Chevalier, Kashyap, and Rossi (2003) and continued by Nevo and Hatzitaskos (2006) in the context of testing the loss leader hypothesis of retail sales. We find that controlling for products with zero demand using our approach gives demand estimates that can be more than twice as elastic than standard estimates that select out the zeroes. We also show that the estimated price elasticities increase substantially during Lent (a high demand period for this product category) after we control for the zeroes. Both of these findings have implications for reconciling the loss-leader hypothesis with the data.

The plan of the paper is the following. In Section 2, we illustrate the stylized empirical pattern of Zipf's law where market zeroes naturally arise. In Section 3, we describe our solution to the zeroes problem using a simple logit setup without random coefficients to make the essential matters transparent. In Section 4, we introduce our general approach for discrete choice model with random coefficients. Section 5 and 6 present results of Monte Carlo simulations and the application to the DFF data, respectively. Section 7 concludes.

2 The Empirical Pattern of Market Zeroes

In this section we highlight some empirical patterns that arise in applications where the zero shares problem arises, which will also help to motivate the general approach we take to it in the paper. Here we will primarily use workhorse store level scanner data to illustrate these patterns. It is this same data that will also be used for our empirical application. However we emphasize that our focus here on scanner data is only for the sake of a concrete illustration of the market zeroes problem - the key patterns we highlight in scanner data are also present in many other economic settings where demand estimation techniques are used (discussed further below and illustrated in the Appendix).

We employ here a widely studied store level scanner data from the Dominick's Finer Foods grocery chain, which is public data that has been used by many researchers.[1] The data comprises 93 Dominick's Finer Foods stores in the Chicago metropolitan area over the years from 1989 to 1997. Like other store level scanner data sets, this data set provides demand information (price, sales, marketing) at store/week/UPC level, where a UPC (universal product code) is a unique bar

[1]For a complete list of papers using this data set, see the website of Dominick's Database: http://research.chicagobooth.edu/marketing/databases/dominicks/index.aspx

code that identifies a product[2].

Table 1 presents information on the resulting product variety across the different product categories in data. The first column shows the number of products in an average store/week - the number of UPC's can be seen varying from roughly 50 (e.g., bath tissue) to over four hundred (e.g., soft drinks) within even these fairly narrowly defined categories. Thus there is considerable product variety in the data. The next two columns illustrate an important aspect of this large product variety: there are often just a few UPC's that dominate each product category whereas most UPC's are not frequently chosen. The second column illustrates this pattern by showing the well known "80/20" rule that prevails in our data: we see that roughly 80 percent of the total quantity purchased in each category is driven by the top 20 percent of the UPC's in the category. In contrast to these "top sellers", the other 80 percent of UPC's contain relatively "sparse sellers" that share the remaining 20 percent of the total volume in the category. The third column shows an important consequence of this sparsity: many UPC's in a given week at a store simply do not sell. In particular, we see that the fraction of observations with zero sales can even be nearly 60% for some categories.

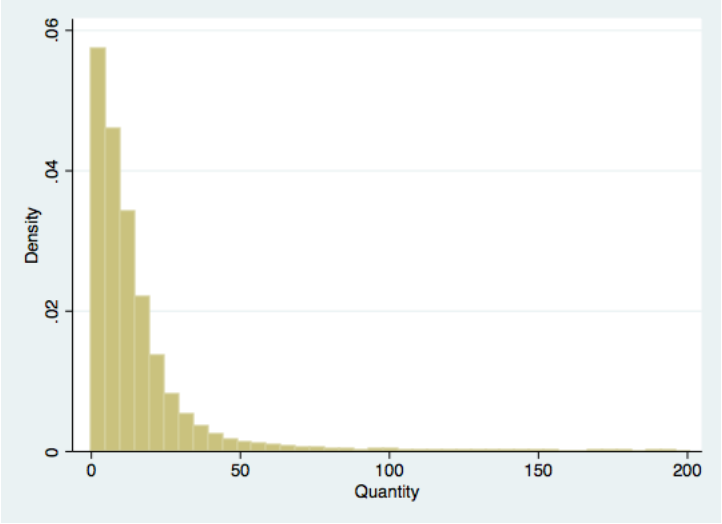
Table 1: Selected Product Categories in the Dominick's Database

Category	Average Number of UPC's in a Store/Week Pair	Percent of Total Sale of the Top 20% UPC's	Percent of Zero Sales
Beer	179	87.18%	50.45%
Cereals	212	72.08%	27.14%
Crackers	112	81.63%	37.33%
Dish Detergent	115	69.04%	42.39%
Frozen Dinners	123	66.53%	38.32%
Frozen Juices	94	75.16%	23.54%
Laundry Detergents	200	65.52%	50.46%
Paper Towels	56	83.56%	48.27%
Refrigerated Juices	91	83.18%	27.83%
Soft Drinks	537	91.21%	38.54%
Snack Crackers	166	76.39%	34.53%
Soaps	140	77.26%	44.39%
Toothbrushes	137	73.69%	58.63%
Canned Tuna	118	82.74%	35.34%
Bathroom Tissues	50	84.06%	28.14%

We can visualize this situation in another way by fixing a product category (here we use canned

[2]Store level scanner data can often be augmented with a panel of household level purchases (available, for example, through IRI or Nielsen). Although the DFF data do not contain this micro level data, the main points of our analysis are equally applicable to the case where household level data is available. In fact our general choice model will accommodate the possibility of micro data. Store level purchase data can be viewed as a special case household level data where all households are observationally identical (no observable individual level characteristics).

Figure 1: Zipf's Law in Scanner Data



tuna) and simply plotting the histogram of the volume sold for each week/UPC realization for a single store in the data. This frequency plot is given in *Figure 1*. As can be see there is a sharp decay in the empirical frequency as the purchase quantity becomes larger, with a long thin tail. In particular the bulk of UPC's in the store have small purchase volume: the median UPC sells less than 10 units a week, which is less than 1.5% of the median volume of Tuna the store sells in a week. The mode of the frequency plot is a zero share.

This power-law decay in the frequency of product demand is often associated with "Zipf's law" or the "the long tail", which has a long history in empirical economics.[3] We present further illustrations of this long-tail demand pattern found in international trade flows as well as cross-county homicide rates in Appendix A, which provides a sense of the generality of these stylized facts.

The key takeaway from these illustrations is that the presence of market zeroes in the data is closely intertwined to the prevalence of power-law patterns of demand. We will exploit this relationship to place structure on the data generating process that underlies market zeroes.

3 A First Pass Through Logit Demand

Why do zero shares create a problem for demand estimation? In this section, we use the workhorse multinomial logit model to explain the zeroes problem and our solution. The general case is treated in the next section.

[3]See Anderson (2006) for a historical summary of Zipf's law and many examples from the social and natural sciences. See Gabaix (1999) for an application of Zipf's law to the economics literature.

3.1 Zeroes Problem: the Logit Case

Consider a multinomial logit model for the demand of J_t products ($j = 1, \ldots, J_t$) and an outside option ($j = 0$). A consumer i derives utility $u_{ijt} = \delta_{jt} + \epsilon_{ijt}$ from product j in market t, where δ_{jt} is the mean-utility of product j in market t, and ϵ_{ijt} is the idiosyncratic taste shock that follows the type-I extreme value distribution. As is standard, the mean-utility δ_{jt} of product $j > 0$ is modeled as

$$\delta_{jt} = x'_{jt}\beta + \xi_{jt}, \tag{3.1}$$

where x_{jt} is the vector of observable (product, market) characteristics, often including price, and ξ_{jt} is the unobserved characteristic. The outside good $j = 0$ has mean utility normalized to $\delta_{0t} = 0$. The parameter of interest is β.

Each consumer chooses the product that yields the highest utility: $s_{ijt} = 1\{u_{jt} \geq u_{j't} \; \forall j' = 0, 1, \ldots, J_t\}$. Aggregating consumers' choices, we obtain the true choice probability of product j in market t, denoted as

$$\pi_{jt} = \Pr(\text{product } j \text{ is chosen in market } t) = E[s_{ijt} | \delta_{1t}, \ldots, \delta_{J_t t}].$$

The standard approach introduced by Berry (1994) for estimating β is to combine demand system inversion and instrumental variables.

First, for demand inversion, one uses the logit structure to find that

$$\delta_{jt} = \log(\pi_{jt}) - \log(\pi_{0t}), \text{ for } j = 1, \ldots, J_t. \tag{3.2}$$

Then, to handle the potential endogeneity of x_{jt} (i.e., its correlation with ξ_{jt}), one finds a random vector z_{jt}, such that

$$E[\xi_{jt} | z_{jt}] = 0. \tag{3.3}$$

Then two stage least squares with δ_{jt} defined in terms of choice probabilities as the dependent variable becomes the identification strategy for β.

Unfortunately π_{jt} is not observed as data - it is a theoretical choice probability defined by the model but only indirectly revealed through actual consumer choices. The standard approach to this following Berry (1994), Berry, Levinsohn, and Pakes (1995), and many subsequent papers in the literature has been to substitute $s_{jt} := n_t^{-1} \sum_{i=1}^{n_t} s_{ijt}$, the empirical market share of product j in market t based on the choices of n_t potential consumers, for π_{jt}, and run a two-stage least square with $\log(s_{jt}) - \log(s_{0t})$ as dependent variable, x_{jt} as covariates, and z_{jt} as instruments to obtain estimates for β.

Plugging in the estimate s_{jt} for π_{jt} appears innocuous at first glance because the number of potential consumers (n) in a market from which s_{jt} is constructed is typically large. Nevertheless problems arise when there are (jt)'s for which π_{jt} is very small. Because the slope of the natural logarithm function approaches infinity when the argument approaches zero, even small estimation error of π_{jt} may lead to large error in the plugged-in version of δ_{jt} when π_{jt} is very small. In par-

ticular, s_{jt} may frequently equal zero in this case, causing the demand inversion to fail completely.

Data sets with zero shares are frequently encountered in empirical research as discussed in the Section 2. With such data, a common practice is to ignore the (jt)'s with $s_{jt} = 0$, effectively lumping those j's into the outside option in market t. This however leads to a selection problem. To see this, suppose $s_{jt} = 0$ for some (j, t) and one drops these observations from the analysis - effectively one is using a selected sample where the selection criterion is $s_{jt} > 0$. In this selected sample, the conditional mean of ξ_{jt} is no longer a constant. This is the well-known selection-on-unobservables problem and with such sample selection, an attenuation bias ensues.[4] The attenuation bias generally leads to demand estimates that appear to be too inelastic.[5]

Another commonly adopted empirical "trick" is to add a small positive number $\epsilon > 0$ to the s_{jt}'s that are zero, and use the resulting modified shares $s^{\epsilon}_{jt} > 0$ in place of π_{jt}. [6] However, this trick only treats the symptom, i.e., $s_{jt} = 0$, but overlooks the nature of the problem: the true choice probability π_{jt} is small. And in this case, small estimation error in any estimator $\widehat{\pi}_{jt}$ of π_{jt} would lead to large error in the plugged-in version of δ_{jt} and the estimation of β. This problem manifests itself directly because the estimate $\widehat{\beta}$ can be incredibly sensitive to the particular choice of the small number being added and there is little guidance on what is the "right" choice of the small number. In general, like selecting away the zeroes, the "adding a small number trick" is also a biased estimator for β. We illustrate both biases in the Monte Carlo section (Section 7).

Despite their failure as general solutions, these "ad hoc zero fixes" have in them what could be a useful idea – Perhaps the variation among the non-zero share observations can be used to estimate the model parameters, while at the same time the presence of zeroes is controlled in such a way that avoids bias. We will present a new estimator that formalizes this possibility by using moment *inequalities* to control for the zeroes in the data while using the variation in the remaining part of the data to consistently estimate the demand parameters. Next we present an asymptotic framework where this intuitive idea can be formalized.

3.2 An Asymptotic Framework Accommodating Zeroes

The existing asymptotic framework for aggregate demand makes assumptions to rule out zeroes in the asymptotic limit as $n_t \to \infty$. For example, Berry, Linton, and Pakes (2004) assume that

[4] To see why $E[\xi_{jt}|x_{jt}, s_{jt} > 0]$ is not a constant, consider two values of x_{jt}: x, x^* such that $x'\beta > x^{*\prime}\beta$, and consider the homoskedastic case for simplicity. For each given value of x_{jt}, the criterion $s_{jt} > 0$ selects high values of ξ_{jt} and leaves out low values of ξ_{jt}. Moreover, the selection is more severe for x^* than for x because the unobservable (to econometricians) needs to more appealing to induce a positive observed market share when the observable characteristic is less appealing.

Thus, we should have

$$E[\xi_{jt}|x_{jt} = x^*, s_{jt} > 0] > E[\xi_{jt}|x_{jt} = x, s_{jt} > 0], \tag{3.4}$$

and clearly, $E[\xi_{jt}|x_{jt}, s_{jt} > 0]$ is not a constant.

[5] It is easy to see that the selection bias is of the same direction if the selection criterion is instead $s_{jt} > 0$ for all t, as one is effectively doing when focusing on a few top sellers that never demonstrate zero sales in the data. The reason is that the event $s_{jt} > 0$ for all t contains the event $s_{jt} > 0$ for a particular t. If the markets (ξ_{jt}'s) are weakly dependent, the particular t part of the selection dominates.

[6] Berry, Linton, and Pakes (2004)and Freyberger (2015) study the biasing effect of plugging in s_{jt} for π_{jt}. Their bias corrections do not apply when there are zeroes in the empirical shares.

$|s_{jt} - \pi_{jt}|/\pi_{jt} \to_p 0$ for logit class models (see Assumption A3 in their paper). This requires $\Pr(s_{jt} = 0) \to 0$ because $\Pr(s_{jt} = 0) \leq \Pr(|s_{jt} - \pi_{jt}|/\pi_{jt} \geq 1)$. Freyberger (2015) directly assumes a fixed lower bound on π_{jt}, which also implies $\Pr(s_{jt} = 0) \to 0$ as $n_t \to \infty$. In their asymptotic limit, $\ln(0)$ is not encountered, the selection problem discussed above disappears. These of course is not realistic in many applications given the significant amount of zeroes that empirical researchers frequently encounter in their data sets with seemingly large n_t.

Assuming that the data $(x_t, s_t, z_t)_{t=1}^T$ come from a many market context ($T \to \infty$, $J_t \leq \bar{J}$ with fixed $\bar{J}$), a realistic asymptotic framework needs to maintain a positive fraction of zeroes in the limit. Letting n_t be fixed as the sample size T grows is one way, but that will prevent the $\log(\pi_{jt}) - \log(\pi_{0t})$ of any product from being revealed in the limit, and thus rule out point identification of the model parameters.[7] A more productive approach is to let $n_t \to \infty$, and to simultaneously allow a non-negligible fraction of products to have π_{jt} drifting to zero at the rate $1/n_t$.

We propose the following simple model. For each product jt, we assume that it is either a safe product in which case, $\pi_{jt} \geq \underline{\varepsilon}_0$ for a positive number $\underline{\varepsilon}_0$, or a risky product, in which case $n_t \pi_{jt} \geq \underline{\varepsilon}_1$ for some positive number $\underline{\varepsilon}_1$. The numbers $\underline{\varepsilon}_0$ and $\underline{\varepsilon}_1$ are not known to the researcher, neither is the identity of the safe products. Since we will use the safe products as the source of identification, we assume that they are characterized by the observable instruments $z_{jt} \in \mathcal{Z}_0$, and $\mathcal{Z}_0$ is a subset of the support of z_{jt} (denoted $\text{supp}(z_{jt})$). Both this subset and the support of z_{jt} may change with t. The set $\mathcal{Z}_0$ is unknown to the researcher. Formally, the conditions are:

Assumption 1 (Safe-Risky Products)**.** *(a) There exists a fixed positive constant $\underline{\varepsilon}_0$ and a sequence $\{\mathcal{Z}_0 \subseteq supp(z_{jt})\}_{t=1,2,\ldots}$, such that for all $t = 1, 2, \ldots$ we have $\inf_{j,t:z_{jt} \in \mathcal{Z}_0} \pi_{jt} \geq \underline{\varepsilon}_0$.*
(b) *For all $t = 1, 2, 3, \ldots$, $\inf_{t=1,\ldots,T} \pi_{0t} \geq \underline{\varepsilon}_0$.*
(c) *There exists a fixed positive constant $\underline{\varepsilon}_1$ such that for all $T = 1, 2 \ldots$, we have $\inf_{j,t:z_{jt} \notin \mathcal{Z}_0} n_t \pi_{jt} \geq \underline{\varepsilon}_1$.*

Note that the outside product is assumed to always be a safe product, which is convenient for theoretical derivations and also quite realistic for virtually all the empirical applications (outside share is typically not close to zero, but often greater than 0.5).

The presence of the risky products with $\pi_{jt} \propto n_t^{-1}$ not only lead to a non-vanishing number of zero shares, but also makes consistently estimating $\delta_{jt} = \log \pi_{jt} - \log \pi_{0t}$ impossible. This is because the best rate at which π_{jt} can be estimated is achieved by its maximum likelihood estimator s_{jt} . This rate is $\sqrt{\pi_{jt}/n_t}$, which is proportional to n_t^{-1}. Plugging an estimator (say $\hat{\pi}_{jt}$) converging at this rate in the logarithm, and we get

$$|\ln \hat{\pi}_{jt} - \ln \pi_{jt}| \geq \frac{1}{\pi_{jt} \vee \hat{\pi}_{jt}} |\hat{\pi}_{jt} - \pi_{jt}| = \frac{1}{O_p(n_t^{-1})} |\hat{\pi}_{jt} - \pi_{jt}| = \frac{n_t |\hat{\pi}_{jt} - \pi_{jt}|}{O_p(1)},$$

which is not $o_p(1)$. Thus, the estimation error of δ_{jt} does not disappear as $n_t \to \infty$ regardless of the ad hoc fixes to s_{jt}. Moreover, since the econometrician typically does not know $\mathcal{Z}_0$, it in general

[7] See the previous version of our paper, Gandhi, Lu, and Shi (2013), for the partial identification approach.

is not possible to discard the risky products without incurring selection bias. In the next section, we propose a novel estimator of model parameters that achieves consistency in this challenging asymptotic framework.

3.3 A Bounds Estimator

Our estimator is based on the same generalized method of moment principle of standard BLP estimators and utilizes the information from the safe products to achieve consistency. However, there are two challenges that we face: (1) the presence of the risky products for which a consistent estimator of δ_{jt} does not exist, and (2) the fact that the identity of the safe products (i.e. $\mathcal{Z}_0$) is unknown. In this section, we describe the estimator first, and then explain its novel features and the roles that they play to overcome the challenges.

The estimator uses two mean-utility estimators δ^u_{jt} and δ^ℓ_{jt}, where $\delta^u_{jt} > \delta^\ell_{jt}$. We refer to them as the upper and lower bounds of δ_{jt} because they bound δ_{jt} from above and below *on average* in the sense discussed in the next subsection. Using these estimators and a countable collection $\mathcal{G}$ of instrumental indicator functions $g : R^{d_z} \to \{0, 1\}$, where d_z is the dimension of z_{jt}, we form the moments

$$\bar{m}^u_T(\beta, g) := T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (\delta^u_{jt} - x'_{jt}\beta) g(z_{jt}) \text{ and}$$
$$\bar{m}^\ell_T(\beta, g) := T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (\delta^\ell_{jt} - x'_{jt}\beta) g(z_{jt}).$$

These moments are used to form the criterion function:

$$\widehat{Q}_T(\beta) = \sum_{g \in \mathcal{G}} \mu(g) \left\{ [\bar{m}^u_T(\beta, g)]^2_- + [\bar{m}^\ell_T(\beta, g)]^2_+ \right\}, \tag{3.5}$$

where $\mu(g) : \mathcal{G} \to [0, 1]$ is a probability mass function on $\mathcal{G}$, $[x]_- = |\min\{0, x\}|$ and $[x]_+ = \max\{0, x\}$. Finally, our parameter estimator $\widehat{\beta}_T$ is the minimizer of $\widehat{Q}_T(\beta)$:

$$\widehat{\beta}_T = \arg\max_{\beta \in B} \widehat{Q}_T(\beta),$$

where B is the parameter space of β.

The key to the consistency of an M-estimator is that the criterion function $\widehat{Q}_T(\beta)$ should be small (close to zero in our case) at the true value β_0, and should be bounded away from zero for β bounded away from β_0, making sure that $\widehat{Q}_T(\beta)$ is minimizes at a point close to β_0 in large samples. Three nonstandard features of our criterion function $\widehat{Q}_T(\beta)$ ensure that it has these properties. First, the bounds δ^u_{jt} and δ^ℓ_{jt} are used instead of a point estimate of δ_{jt}. Second, the moments enter the criterion function through a negative part and a positive part function. And third, a countable collection of indicator functions of z_{jt} is employed to form moment conditions,

instead of a finite number of full support instrumental functions of z_{jt}. The requirement for this collection will be similar to that for the collection of instruments in Andrews and Shi (2013). We explain how the three features work together now.

First, as described in the next subsection, the bounds δ_{jt}^u and δ_{jt}^ℓ will be constructed to satisfy:

$$\begin{aligned}
&\Pr\left(T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt}^u g(z_{jt}) \geq T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt} g(z_{jt}) - c\right) \to 1 \\
&\Pr\left(T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt}^\ell g(z_{jt}) \leq T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt} g(z_{jt}) + c\right) \to 1,
\end{aligned} \tag{3.6}$$

for arbitrarily small $c > 0$, as $T \to \infty$, and for any bounded nonnegative-valued function $g(\cdot)$ of z_{jt}. This combined with the second feature–$\widehat{Q}_T(\beta)$ only responds to the negative part of $\bar{m}_T^u(\beta, g)$ and to the positive part of $\bar{m}_T^\ell(\beta, g)$–implies that the criterion function is small when evaluated the true value.

Second, the construction of the bounds δ_{jt}^u and δ_{jt}^ℓ below will also ensure that they collapse to each other and to δ_{jt} asymptotically for the safe products ($z_{jt} \in \mathcal{Z}_0$). Then, with any g such that $g(z) = 0,\ \forall z \notin \mathcal{Z}_0$, we have

$$T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt}^u g(z_{jt}) = T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt}^\ell g(z_{jt}) + o_p(1) = T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt} g(z_{jt}) + o_p(1). \tag{3.7}$$

Let $\mathcal{G}_0$ denote the subset of $\mathcal{G}$ containing the g's with support lying in $\mathcal{Z}_0$, i.e.,

$$\mathcal{G}_0 = \{g \in \mathcal{G} :\ g(z) = 0\ \forall z \notin \mathcal{Z}_0\}.$$

Then the part of $\widehat{Q}_T(\beta)$ with $g \in \mathcal{G}_0$ is

$$\begin{aligned}
\widehat{Q}_T^0(\beta) &:= \sum_{g\in\mathcal{G}_0}\mu(g)\left\{[\bar{m}_T^u(\beta, g)]_-^2 + [\bar{m}_T^\ell(\beta, g)]_+^2\right\} \\
&= o_p(1) + \sum_{g\in\mathcal{G}}\mu(g)\left(T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt} - x_{jt}'\beta)g(z_{jt})\right)^2.
\end{aligned}$$

This part behaves as a GMM criterion function where δ_{jt} is used directly and it is bounded away from zero for β bounded away from β_0, if $\mathcal{Z}_0$ is rich enough and $\mathcal{G}_0$ contains enough number of functions on $\mathcal{Z}_0$. The richness of $\mathcal{Z}_0$ will be imposed as a rank condition which is the standard BLP identification condition imposed on the safe products only.

Third, the richness of $\mathcal{G}_0$ is ensured by the construction of $\mathcal{G}$ – a task that is standard if $\mathcal{Z}_0$ is known and less so in our case where $\mathcal{Z}_0$ is unknown. Our idea comes from the construction of instruments for moment inequality models in Andrews and Shi (2013). From Andrews and Shi (2013), if $\mathcal{G}_0$ contains all indicator functions of hypercubes $B_g \subseteq \mathcal{Z}_0$, it is rich enough to preserve

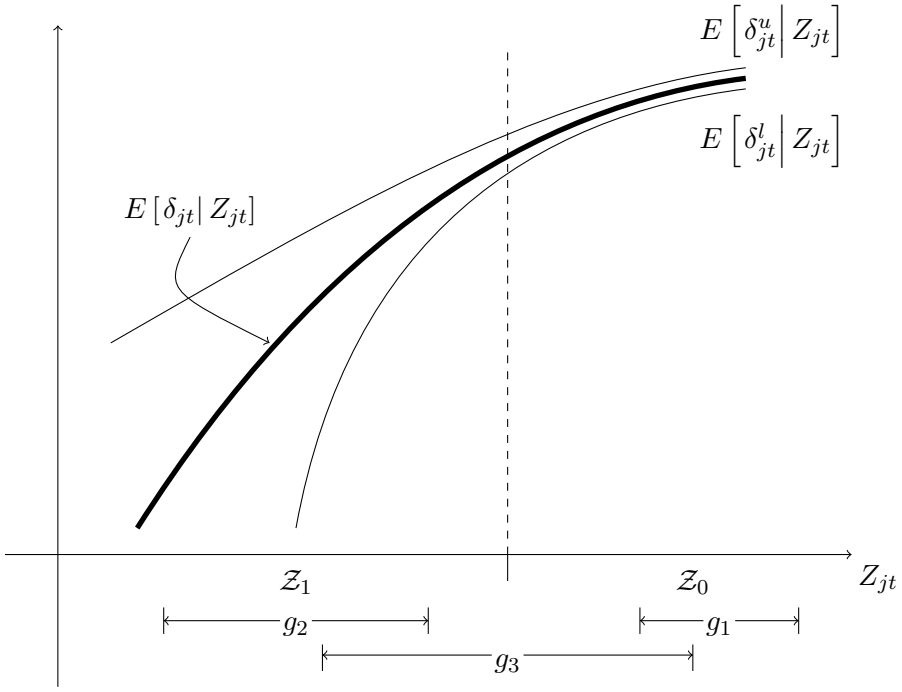
the identification information provided by the richness of $\mathcal{Z}_0$. With $\mathcal{Z}_0$ unknown, a simple way to ensure that is to let $\mathcal{G}$ contain all indicator functions of hypercubes in $\text{supp}(z_{jt})$. Andrews and Shi (2013) also show that a countable reduction of the set of all indicators of hypercubes works just as well and is easier to implement in practice. Therefore, that is the choice that we make for $\mathcal{G}$, which we describe in detail in Section 4.1 below. Finally, it is important to emphasize that identification may not be achieved if one only uses instrumental functions that have global support (i.e. support on the entire $\text{supp}\,(z_{jt})$). This is because (3.7) does not hold for g's with global support, and therefore $\widehat{Q}_T(\beta)$ may not be bounded away from zero for β bounded away from β_0.

To gain more intuition of the above arguments, assume that the sample moments converge to the population expectation (which is not needed for our formal arguments below). Then the expectation version of (3.6) is

$$E[\delta^u_{jt}|z_{jt}] \geq E[\delta_{jt}|z_{jt}] \geq E[\delta^\ell_{jt}|z_{jt}].$$

For $z_{jt} \in \mathcal{Z}_0$, the two bounds collapse, while for $z_{jt} \in \mathcal{Z}_1$, the bounds may remain slack. *Figure 2* provides a graphical illustration of the above arguments. In the safe products region $\mathcal{Z}_0$, the bounds are tight and provide identification power, while in $\mathcal{Z}_1$, the bounds may be uninformative but still valid. So instrumental functions such as $g_1 \in \mathcal{G}_0$ will form moment equalities that point identify the model. Other instrumental functions, such as $g_2, g_3 \in \mathcal{G}_1 \equiv \mathcal{G}\backslash\mathcal{G}_0$, are associated with slack moment inequalities which do not contribute to but also do not undermine identification.

Figure 2: Illustration of Bounds Approach



3.4 Construction of the Bounds

Next we describe the construction of δ_{jt}^u and δ_{jt}^ℓ for the logit case. Recall that $\delta_{jt} = \log(\pi_{jt}) - \log(\pi_{0t})$. The piece $\log(\pi_{0t})$ in δ_{jt} is not problematic because π_{0t} is bounded away from zero under Assumption 1(b). We can plug in s_{0t} or any modification $\tilde{s}_{0t}$ of s_{0t} for π_{0t}. As long as the modification is negligible relative to the estimation error in s_{0t}, standard arguments will imply

$$T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}[\log(\tilde{s}_{0t}) - \log(\pi_{0t})] = o_p(1). \tag{3.8}$$

On the other hand, the piece $\log(\pi_{jt})$ is potentially difficult to approximate because π_{jt} can be close to zero: simply plugging in s_{jt} for π_{jt} is problematic because s_{jt} can be zero or very small. Instead, we propose bounding $\log(\pi_{jt})$ with:

$$\log((n_t s_{jt} + \iota_u)/n_t) \text{ and } \log((n_t s_{jt} + \iota_\ell)/n_t), \tag{3.9}$$

where ι_u and ι_ℓ are two positive numbers to be determine numerically.

To determine ι_u and ι_ℓ, note that $n_t s_{jt}$ follows a binomial distribution: $Bin(n_t, \pi_{jt})$ [8]. For each fixed $n_t = n$, $\pi_{jt} = \pi$, and $\iota \geq 0$, define the function

$$f(\iota; n, n\pi) := E[\log(ns_{jt} + \iota) - \log(n\pi)].$$

The function f is negative infinity at $\iota = 0$ (because s_{jt} can be 0 with a positive probability), strictly increasing with ι, and approaches positive infinity as $\iota \to \infty$. Therefore, at each n and $n\pi$, the function crosses zero once and only once. The point of crossing $\iota^*(n, n\pi)$ can be numerically calculated because the function $f(\iota; n, n\pi)$ (i.e., the expectation) can be calculated using the binomial distribution. We can find any finite fixed numbers ι_u and ι_ℓ that satisfy

$$\iota_u \geq \underline{\iota}_u := \sup_{n,\pi: n\pi \geq \underline{\varepsilon}_1} \iota^*(n, n\pi) \text{ and } 0 \leq \iota_\ell \leq \bar{\iota}_\ell := \inf_{n,\pi: n\pi \geq \underline{\varepsilon}_1} \iota^*(n, n\pi). \tag{3.10}$$

Such a construction immediately yields:

$$E[\log((n_t s_{jt} + \iota_u)/n_t) - \log(\pi_{jt})|z_{jt}] \geq 0 \text{ and } E[\log((n_t s_{jt} + \iota_\ell)/n_t) - \log(\pi_{jt})|z_{jt}] \leq 0. \tag{3.11}$$

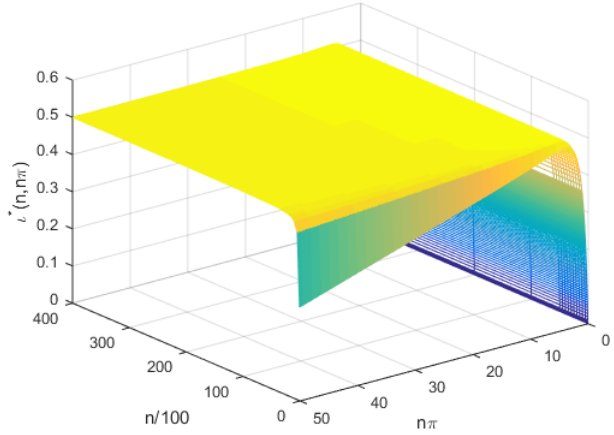
This and (3.8), together with appropriate law of large number will imply (3.6), when we define

$$\begin{aligned} \delta_{jt}^u &= \log((n_t s_{jt} + \iota_u)/n_t) - \log(\tilde{s}_{0t}) \\ \delta_{jt}^\ell &= \log((n_t s_{jt} + \iota_\ell)/n_t) - \log(\tilde{s}_{0t}). \end{aligned} \tag{3.12}$$

We specify $\tilde{s}_{0t}$ in the general case later. For the logit case, $\tilde{s}_{0t} = s_{0t}$ works just fine.

[8] Here we maintain the standard assumption that in each given market, consumers' choices are independent and identically distributed.

Figure 3: $\iota^*(n, n\pi)$ for a Range of n and $n\pi$ Values



To see why the bounds also satisfy (3.7), we need to examine the magnitude of $\underline{\iota}_u$ and $\bar{\iota}_\ell$. Our calculation of $\iota^*(n, n\pi)$ for a large range of n and $n\pi$ is shown in Figure 3. It shows that $\underline{\iota}_u \approx 0.51$ and it does not depend on the lower bound for $n\pi$: $\underline{\varepsilon}_1$, and $\bar{\iota}_\ell > 0$ as long as $\underline{\varepsilon}_1 > 0$. Importantly, both $\underline{\iota}_u$ and $\bar{\iota}_\ell$ are fixed finite numbers, and thus ι_u and ι_ℓ can be any fixed finite numbers from the range defined in (3.10). Thus, $(n_t s_{jt} + \iota_u)/n_t$ and $(n_t s_{jt} + \iota_\ell)/n_t$ inherit all the nice convergent properties of s_{jt} for the safe products $\pi_{jt} > \underline{\varepsilon}_0$. Thus, (3.7) follows from standard asymptotic arguments.

4 The General Model and Estimator

Now we extend the bound construction to the general differentiated product demand model and present our parameter estimator. The specification of the general model is the same as the logit model except that the consumer level shock ϵ_{ijt} in $u_{ijt} = \delta_{jt} + \epsilon_{ijt} \equiv x'_{jt}\beta + \xi_{jt} + \epsilon_{ijt}$ is no longer type-I extreme value distribution. Instead, we assume that

$$\epsilon_{it} = (\epsilon_{i0t}, \ldots, \epsilon_{iJ_tt}) \sim F\left(\cdot \,|x_t; \lambda\right), \tag{4.1}$$

where x_t stands for $(x'_{1t}, \ldots, x'_{J_tt})'$, and $F(\cdot|x_t, \lambda)$ is a conditional cumulative distribution function known up to the finite dimensional unknown parameter λ. By allowing x_t and an unknown parameter to enter the distribution of ϵ_{ijt}, this specification is general enough to encompass most models used in empirical work. In particular, it encompasses the random coefficient specifications $\epsilon_{ijt} = x'_{jt}(\beta_i - \beta) + \nu_{ijt}$, where β_i is a vector of random coefficients that follows a distribution (e.g., joint normal) known up to some unknown parameter, ν_{ijt} is the idiosyncratic taste shock.[9]

Given the specification, the unknown parameter in the general model is $\theta = (\beta', \lambda')'$. For clarity,

[9] Requiring $F(\cdot|x_t, \lambda)$ to be known up to a finite dimensional parameter rules out the vertical model (see Berry and Pakes (2007)) because for the vertical model, ϵ_{ijt} is a function of the unobservable product characteristics (quality).

we use $\theta_0 \equiv (\beta_0', \lambda_0')'$ to denote the true value of θ. Let $B \subseteq R^{d_\beta}$ denote the parameter space of β, and $\Lambda \subseteq R^{d_\lambda}$ the parameter space of λ. Let $\Theta = B \times \Lambda$ be the parameter space of θ.

In the general model, the choice probability of each product is determined by:

$$\pi_{jt} = \int 1\{\delta_{jt} + e_j \geq \max_{j'=0,1,\ldots,J_t} (\delta_{j't} + e_{j'})\} dF(e_1, \ldots, e_{J_t} | x_t, \lambda_0), \ j = 1, \ldots, J_t. \tag{4.2}$$

This system is invertible under the connected substitute condition in Berry, Gandhi, and Haile (2013). In other words, we can define the inverse demand function $\delta_t(\pi_t, \lambda) := (\delta_{jt}(\pi_t, \lambda))_{j=1}^{J_t}$ as the solution to the equation system

$$\pi_{jt} = \int 1\{\delta_{jt}(\pi_t, \lambda) + e_j \geq \max_{j'=0,1,\ldots,J_t} \left(\delta_{j't}(\pi_t, \lambda) + e_{j'}\right)\} dF(e_1, \ldots, e_{J_t} | x_t, \lambda), \ j = 1, \ldots, J_t. \tag{4.3}$$

Inverting the demand system allows for the use of instrumental variables to identify θ based on the exclusion restriction:

$$E\left[\xi_{jt} \,|z_{jt}\right] = 0. \tag{4.4}$$

This is because one can then obtain the following moment restriction:

$$E\left[\delta_{jt}(\pi_t, \lambda_0) - x_{jt}'\beta_0 \middle| \, z_{jt}\right] = 0. \tag{4.5}$$

If π_t were observed, the parameters (λ and β) in the model would be identified under standard GMM identification conditions. However, as discussed in the logit case, π_t is not observed. Instead only a noisy measure s_t is, and s_t frequently contains zero elements when π_t is not bounded away from the boundary of the probability simplex. As in the logit case, $\delta_t(s_t, \lambda)$ is typically not well defined when s_t contains zero elements, and thus simply substituting s_t for π_t in the moment conditions (4.5) is problematic.

As in the logit case, we seek to identify and estimate the model parameters based on the safe products, while controlling for the effect of the risky products. We adopt the asymptotic framework laid out in Section 3.2 and impose Assumption 1.

4.1 Bound Estimator for the General Case

Like in the logit case, we construct a pair of inverse demand functions: $\delta_{jt}^u(s_t, \lambda)$ and $\delta_{jt}^\ell(s_t, \lambda)$, to form bounds for $\delta_{jt}(\pi_t, \lambda)$. The construction follows from the logit case but adjust for the different functional form:

$$\begin{aligned}
\delta_{jt}^u(s_t, \lambda) &= \log((n_t s_{jt} + \iota_u)/n_t) + \delta_{jt}(\tilde{s}_t, \lambda) - \log(\tilde{s}_{jt}), \\
\delta_{jt}^\ell(s_t, \lambda) &= \log((n_t s_{jt} + \iota_\ell)/n_t) + \delta_{jt}(\tilde{s}_t, \lambda) - \log(\tilde{s}_{jt}),
\end{aligned} \tag{4.6}$$

where ι_ℓ and ι_u are fixed numbers, and $\tilde{s}_t$ is a slight modification of s_t to take it off the boundary of the probability simplex (e.g., the Laplace shares). The formal requirements on ι_ℓ, ι_u and $\tilde{s}_t$ involes

technical details and will be discussed in Sections 4.2 and 5.

With the bounds defined in (4.6), we can extend (3.5) to define our estimator for θ in the general case:

$$\widehat{\theta}_T := (\widehat{\beta}_T', \widehat{\lambda}_T')' = \arg\min_{\theta \in \Theta} \widehat{Q}_T(\theta), \tag{4.7}$$

where

$$\widehat{Q}_T(\theta) = \sum_{g \in \mathcal{G}} \mu(g) \left\{ [\bar{m}_T^u(\theta, g)]_-^2 + [\bar{m}_T^\ell(\theta, g)]_+^2 \right\}, \tag{4.8}$$

$$\bar{m}_T^u(\theta, g) := T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (\delta_{jt}^u(s_t, \lambda) - x_{jt}'\beta) g(z_{jt}) \text{ and}$$
$$\bar{m}_T^\ell(\theta, g) := T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (\delta_{jt}^\ell(s_t, \lambda) - x_{jt}'\beta) g(z_{jt}).$$

where $\mu(g) : \mathcal{G} \to [0, 1]$ is a probability mass function on $\mathcal{G}$, $[x]_- = \min\{0, x\}$ and $[x]_+ = \max\{0, x\}$, and $\mathcal{G}$ is a collection of instrumental functions.

4.2 Some Implementation Details

To implement the estimator (4.7), we need to specify ι_ℓ, ι_u, $\mathcal{G}$, and $\mu(\cdot)$. In the following, we provide some practical guidance on specifying them based on our experiences in the simulations and the empirical application of this paper.

Since setting $\iota_\ell = 0$ causes numerical breakdown, we let ι_ℓ be a numerical infinitesimal positive number[10]. For ι_u, we set it to 2. We shall see that these choices satisfy the requirements in (3.10) for the logit case and we shall see that they are also valid for the general case (see the formal assumptions in Section 5). Also, it is worth mentioning that in small samples ($T \leq 100$), one might see mild bias in the estimator due to the bounds not being slack enough for the risky products; making the bounds more slack by slightly increasing ι_u reduces the bias without affecting the variance. This is because such changes of ι_u make the bounds more slack for the risky products and thus reducing their chance of biasing the estimates, but only has negligible effect on the moment functions for the safe products because their market shares are much larger than ι_u/n_t.

For $\mathcal{G}$, we divide the instrument vector z_{jt} into discrete instruments, $z_{d,jt}$, and continuous instruments $z_{c,jt}$. Without loss of generality assume that $z_{c,jt}$ lies in $[0, 1]^{d_{z_c}}$.[11] Let the set $\mathcal{Z}_d$ be the discrete set of values that $z_{d,jt}$ can take. The set $\mathcal{G}$ is defined as

$$\mathcal{G} = \{g_{a,r,\zeta}(z_d, z_c) = 1((z_c', z_d')' \in C_{a,r,\zeta}) : C_{a,r,\zeta} \in \mathcal{C}\}, \text{ where}$$

[10] In Matlab, we use the floating-point accuracy "eps" (or 2^{-52}) as ι_ℓ.

[11] If not, we can normalize it to lie in $[0, 1]$ as suggested in Andrews and Shi (2013). For example, we can let $\tilde{z}_{c,jt} = F_{N(0,1)}\left(\widehat{\Sigma}_{z_c}^{-1/2} z_{c,jt}\right)$, where $F_{N(0,1)}(\cdot)$ is the standard normal cdf and $\widehat{\Sigma}_{z_c}$ is the sample covariance matrix of $z_{c,jt}$, and use $\tilde{z}_{c,jt}$ in place of $z_{c,jt}$ to construct the instrumental functions.

$$\mathcal{C} = \{(\times_{u=1}^{d_{z_c}}((a_u-1)/(2r), a_u/(2r)]) \times \{\zeta\} : \ a_u \in \{1,2,...,2r\}, \text{ for } u = 1,...,d_{z_c},$$
$$r = r_0, r_0+1, ..., \text{ and } \zeta \in \mathcal{Z}_d\}. \tag{4.9}$$

In practice, we truncate r at a finite value $\bar{r}_T$. This does not affect the first order asymptotic property of our estimator as long as $\bar{r}_T \to \infty$ as $T \to \infty$. For $\mu(\cdot)$, we use

$$\mu(\{g_{a,r,\zeta}\}) \propto (100+r)^{-2}(2r)^{-d_{z_c}} K_d^{-1}, \tag{4.10}$$

where K_d is the number of elements in $\mathcal{Z}_d$. The same μ measure is used and works well in Andrews and Shi (2013).[12]

5 Consistency

In this section, we establish the consistency for the estimator defined in (4.7). First of all, we impose a high-level assumption to ensure that the bounds (evaluated at the true value) are valid asymptotically over the entire support of z_{jt}, which is analogous to (3.6) for the logit case, and that for the safe products, the bounds collapse to each other, which is analogous to (3.7) for the logit case. The primitive conditions and verification of this high-level assumption will be discussed in detail in Subsection 5.1.

Assumption 2. (a) $\sum_{g\in\mathcal{G}} \mu(g)[\bar{m}_T^u(\theta_0,g)]_-^2 = o_p(1)$ *and* $\sum_{g\in\mathcal{G}} \mu(g)[\bar{m}_T^\ell(\theta_0,g)]_+^2 = o_p(1)$.

(b) $\sup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_0} |\bar{m}_T^u(\theta,g) - \bar{m}^T(\theta,g)| = o_p(1)$ *and* $\sup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_0} |\bar{m}_T^\ell(\theta,g) - \bar{m}^T(\theta,g)| = o_p(1)$

An immediate implication of Assumption 2 is that the objective function (4.8) converges to zero at the true value, i.e., $\widehat{Q}_T(\theta_0) = o_p(1)$, and behaves as an infeasible standard BLP criterion function with known $\mathcal{Z}_0$ at all points in Θ. To ensure consistency, we would need the infeasible criterion function to point identify θ_0. The next assumption, formalizing the intuition discussed in the logit case (Subsection 3.3), leverages on the safe products to identify the model. In particular, it requires that the safe products provide enough information for the identification of the true parameter.

Assumption 3. *For any $c > 0$, there exists $C(c) > 0$ such that*

$$\lim_{T\to\infty} \Pr\left(\inf_{\theta\in\Theta:\|\theta-\theta_0\|>c} \widehat{Q}_T^*(\theta) > C(c)\right) = 1,$$

where $\widehat{Q}_T^*(\theta) = \sum_{g\in\mathcal{G}_0} \mu(g)\bar{m}_T(\theta,g)^2$, with $\bar{m}_T(\theta,g) = T^{-1}\sum_{t=1}^T \sum_{j=1}^{J_t}(\delta_{jt}(\pi_t,\lambda) - x_{jt}'\beta)g(z_{jt})$.

Remark. The function $\widehat{Q}_T^*(\theta)$ is not a feasible criterion function to use because $\mathcal{Z}_0$ is unknown. In our estimator, the identification information contained in $\widehat{Q}_T^*(\theta)$ is automatically utilized because our construction of the bounds guarantees that $\widehat{Q}_T^*(\theta)$ provide an asymptotic lower bound for our criterion function $\widehat{Q}_T(\theta)$ at all $\theta \in \Theta$, and is asymptotically the same as $\widehat{Q}_T(\theta)$ at $\theta = \theta_0$.

[12] Note that appropriate choices of $\mathcal{G}$ and μ are not unique. For other possible choices, see Andrews and Shi (2013).

The following theorem shows the consistency of the bound estimator.

Theorem 1. *Suppose that Assumptions 2 and 3 hold. Then* $\|\widehat{\theta}_T - \theta_0\| \to_p 0$.

5.1 Verification of Assumption 2

In this subsection, we verify Assumption 2. We shall provide primitive conditions on ι_ℓ, ι_u and $\tilde{s}_t$, which, combining with some more standard assumptions given in Assumption 7 below, imply the high-level condition Assumption 2.

First of all, we introduce some basic assumptions on the modified share $\tilde{s}_t$ and the number of consumers n_t:

Assumption 4. (a) $\sup_{t=1,\ldots,T} n_t\|\tilde{s}_t - s_t\| = O_p(1)$ *as* $T \to \infty$.

(b) $\log(\overline{n}_T)/\sqrt{T} \to 0$ *and* $\underline{n}_T \to \infty$, *where* $\overline{n}_T = \max_{t=1,\ldots,T} n_t$ *and* $\underline{n}_T = \min_{t=1,\ldots,T} n_t$.

Part (b) of the above assumption requires n_t to be not too big. This is needed to guarantee a uniform convergence which requires $\log(\iota_u/n_t)$ to not increase too quickly with T. It is a weak assumption since it allows n_t to increase almost exponentially with $\sqrt{T}$.

Next, we discuss the requirements on ι_ℓ and ι_u. The requirements depend on the demand model used. To begin, consider the upper bound

$$\begin{aligned}\delta^u_{jt}(s_t,\lambda) - \delta_{jt}(\pi_{jt},\lambda) = {} & [\log((n_t s_{jt} + \iota_u)/n_t) - \log(\pi_{jt})] \\ & + [(\delta_{jt}(\tilde{s}_t,\lambda) - \log(\tilde{s}_{jt})) - (\delta_{jt}(\pi_{jt},\lambda) - \log(\pi_{jt}))].\end{aligned}$$

We already know from the logit case that the first summand is nonnegative in expectation conditional on π_{jt} as long as $\iota_u \geq 0.51$. The bound $\delta^u_{jt}(s_t,\lambda)$ will be asymptotically valid if either (i) the second summand is asymptotically negligible, or (ii) the conditional expectation of the second summand can be bounded from above by that of the first with an appropriate choice of ι_u. The first case applies to logit-based models, while the second case applies to models where the idiosyncratic error has a thinner tail than the logistic distribution, for example, normal distributions. We discuss them separately next and give examples for each case.

When $\delta_{jt}(\cdot,\lambda) - \log(\cdot_j)$ is Uniformly Continuous

Let $\Delta^0_{J_t}$ denote a subset of $\{\pi \in (0,1)^{J_t} : 1 - \mathbf{1}'_{J_t}\pi \geq \underline{\varepsilon}_0\}$ that π_t can take value in. Let $\Delta^1_{J_t}$ denote an $\underline{\varepsilon}_0/2-$expansion of $\Delta^0_{J_t}$, that is, $\Delta^1_{J_t} = \{\pi \in (0,1)^{J_t} : \pi'\mathbf{1}_{J_t} < 1, \min_{p\in\Delta^0_{J_t}} \|p - \pi\| \leq \underline{\varepsilon}_0/2\}$. Define the function $\hat{\delta}_t(\cdot,\lambda) = (\hat{\delta}_{1t}(\cdot,\lambda),\ldots,\hat{\delta}_{J_t t}(\cdot,\lambda))' : \Delta^0_{J_t} \to R^{J_t}$ where

$$\hat{\delta}_{jt}(\pi,\lambda) := \delta_{jt}(\pi,\lambda) - \log(\pi_j).$$

Since $\Delta^1_{J_t}$ (as well as $\Delta^0_{J_t}$) may contain points arbitrarily close to the boundary of the probability simplex, in general neither $\delta_{jt}(\cdot,\lambda)$ nor $\log(\cdot_j)$ is uniformly continuous on $\Delta^1_{J_t}$. Thus, neither $\delta_{jt}(\tilde{s}_t,\lambda) - \delta_{jt}(\pi_t,\lambda)$ or $\log(\tilde{s}_{jt}) - \log(\pi_{jt})$ may converge to zero as $n_t \to \infty$ and $\pi_{jt} \to 0$ even if $\tilde{s}_t$ is

the most efficient consistent estimate of π_t. However, in many models used in empirical work the logit inverse demand $(\log(\pi_j) - \log(\pi_0))$ is a good first-order approximation of $\delta_{jt}(\pi, \lambda)$ when π_j is close to zero and this first order term is the entire reason that the inverse demand is not uniformly continuous. For such models, it is reasonable to require the following Assumption 5.

Assumption 5. (a) $\max_{t=1,\ldots,T} \sup_{\pi_t, \hat{\pi}_t \in \Delta^1_{J_t} : \pi_t \neq \hat{\pi}_t} \sup_{\lambda \in \Lambda} \frac{\|\hat{\delta}_t(\hat{\pi}_t, \lambda) - \hat{\delta}_t(\pi_t, \lambda)\|}{\|\hat{\pi}_t - \pi_t\|} \leq O_p(1)$,
(b) $0 \leq \iota_\ell \leq \bar{\iota}_\ell$, $\bar{\iota}_\ell > 0$, and $0 < \underline{\iota}_u \leq \iota_u < \infty$, *and*

Now we give a few examples, where Assumption 5(a) is satisfied.

Example 1. Nested Logit. The inverse demand of the nested logit model can be written as $\delta_{jt}(\pi_t, \lambda) = \log(\pi_{jt}/\pi_{0t}) - \theta \log(\pi_{gt}/\pi_{0t})$, where $\pi_{0t} = 1 - \mathbf{1}'_{J_t} \pi_t$ is the outside share and π_{gt} is the aggregate share of all the products in the nest that j is in. In this case, $\hat{\delta}_{jt}(\pi_t, \lambda) = (\theta - 1) \log \pi_{0t} - \theta \log(\pi_{gt})$. Assumption 5(a) is satisfied as long as π_{0t} and π_{gt} are bounded away from zero.

Example 2. Random Coefficient Logit. For the random coefficient logit model, $\delta_{jt}(\pi_t; \lambda)$ is the solution to the following equation system:

$$\pi_{jt} = \exp(\delta_{jt}) \int \frac{\exp(w'_{jt} v)}{1 + \sum_{k=1}^{J_t} \exp(\delta_{kt} + w'_{kt} v)} dF(v; \lambda), \quad j = 1, \ldots, J_t,$$

where w_{jt} is a vector of covariates with random coefficients, and $F(\cdot; \lambda)$ is the distribution of the random coefficient known up to the unknown parameter λ. Using the definition of $\hat{\delta}_{jt}$ above, we can write

$$\exp(-\hat{\delta}_{jt}(\pi_t; \lambda)) = \int \frac{\exp(w'_{jt} v)}{1 + \sum_{k=1}^{J_t} \exp(\hat{\delta}_{kt}(\pi_t; \lambda) + w'_{kt} v) \pi_{kt}} dF(v; \lambda). \tag{5.1}$$

Assume that $\|w_{jt}\|$ is bounded by $\overline{w}$ and $0 < \sup_{\|w\| \leq \overline{w}} \int \exp(w'v) dF(v; \lambda) < \infty$. We can already see that $\hat{\delta}_{jt}(\pi_t; \lambda)$ is bounded away from $-\infty$ when $\pi_{jt} \to 0$ (in which case, $\delta_{jt}(\pi_t; \lambda) \to -\infty$). With additional algebra, we can show that $\partial \hat{\delta}_{jt}(\pi_t; \lambda)/\partial \pi_t$ is bounded, which essentially guarantees Assumption 5(a). The details are given in Appendix C.

With Assumption 5(a) imposed, all we need for ι_u is $\iota_u \geq \sup_{n, n\pi \geq \underline{\varepsilon}_1} \iota^*(n, n\pi)$ which is imposed in Assumption 5(b).

When $\delta_{jt}(\cdot, \lambda) - \log(\cdot_j)$ is Not Uniformly Continuous

In some models used in empirical work, Assumption 5 can fail to hold. For example, if the model is a simple probit with $J_t = 1$, $\delta_t(\pi) = \Phi^{-1}(\pi)$, where Φ^{-1} is the inverse of the standard normal cdf. In this case, $\delta_t(\pi) - \log(\pi) = \Phi^{-1}(\pi) - \log \pi$. This function approaches infinity when $\pi \to 0$, and has arbitrarily large slope near zero. For such cases, an alternative assumption may be reasonable and this is given in parts (a)-(b) of the following Assumption. Part (c) imposes mild further restrictions on $\tilde{s}_{jt}$, ι_ℓ, and ι_u. Note that ι_ℓ is allowed to be zero.

Assumption 6. (a) $\max_{j,t} E[\hat{\delta}_{jt}(\tilde{s}_t, \lambda_0) - \hat{\delta}_{jt}(\pi_t, \lambda_0)|\pi_t, z_t] \leq 0$,

(b) $\min_{j,t} E[\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0)|\pi_t, z_t] \geq 0$,

(c) $\sup_j \sup_{\pi:\pi_j \geq (\underline{\varepsilon}_1 \wedge 1)/n_t} \delta_{jt}(\pi, \lambda_0) \leq C_0 \log(n_t)$ *for a constant* $C_0 > 0$ *for all* t, *and*

(d) *for* $j = 1, \ldots, J_t$, $\tilde{s}_{jt} = s_{jt} + 1/n_t$, $0 \leq \iota_\ell \leq \bar{\iota}_\ell$, $0 < \bar{\iota}_\ell < 1$ *and* $1 < \iota_u < \infty$.

Example 3. Binary Probit. For binary probit model, we can verify by simulation that parts (a)-(b) hold given that (d) holds. Part (c) holds simply because of the shape of $\Phi^{-1}(\cdot)$ which increases slower than $\log(\cdot)$ as the argument decreases to zero.

The next set of assumptions are standard as in Freyberger (2015).

Assumption 7. (a) *The equation system* (4.3) *uniquely defines* $\delta_t(\pi_t, \lambda)$ *for all* t, *all* $\pi_t \in \Delta_{J_t}$ *and all* $\lambda \in \Lambda$.

(b) *In each market, consumers' preferences* $(\epsilon_{ijt})_{j=1}^{J_t}$ *are i.i.d. draws from the known distribution* $F(\cdot\,|x_t; \lambda_0)$ with unknown parameter $\lambda_0 \in \Lambda$. *Consumer choice is determined by* (4.2).

(c) *The moment condition* (4.5) *holds.*

(d) $J_t \leq \bar{J}$ *for all* t *for a fixed integer* $\bar{J}$.

(e) $(x_t, s_t, z_t)_{t=1}^T$ *are independent across market.*

(f) *There exists a constant* M *such that* $E[\xi_{jt}^{2+c}] < M$ *for all* $j = 1, \ldots, J_t$, *all* $t = 1, \ldots, T$, *and all* T *for some* $c > 0$.

(g) $\max_{t=1,\ldots,T} \max_{j=1,\ldots,J_t} |s_{jt} - \pi_{jt}| \to_p 0$ *as* $T \to \infty$.

Finally, we also need a uniform Lipschitz continuity assumption on the function $\delta_{jt}(\cdot, \lambda)$ in order to show that the bounds collapse for the safe products. Note that this assumption has a similar form as Assumption 5(a) above, but is imposed on $\delta_{jt}(\cdot, \lambda)$ instead of $\hat{\delta}_{jt}(\cdot, \lambda)$ and is restricted to a small neighborhood around the safe-products.

Assumption 8. $\sup_{t=1,\ldots,T} \sup_{\lambda \in \Lambda} \sup_{\pi_t, \hat{\pi}_t \in \Delta^1_{J_t}: \pi \neq \hat{\pi}, \pi_j, \hat{\pi}_j \geq \underline{\varepsilon}_0/2} \frac{|\delta_{jt}(\hat{\pi}, \lambda) - \delta_{jt}(\pi, \lambda)|}{\|\hat{\pi} - \pi\|} \leq O_p(1)$.

The following theorem verifies Assumption 2 and its proof can be found in Appendix B.2.

Theorem 2. *Suppose that Assumptions* 1, 4, *and* 7-8 *hold.*

(i) *Then Assumption* 2(b) *holds.*

(ii) *If in addition either Assumption* 5 *or Assumption* 6 *holds, then Assumption* 2(a) *also holds.*

5.2 Partial Identification as an Alternative

The approach above provides a consistent point estimator based on an underlying set of moment inequalities. Point estimation relies on assumptions given above that allow our estimator to automatically use the variation among safe products to ensure consistency. Those assumptions are natural in many applications where the long tail pattern is present and we illustrate its performance in the Monte Carlo below. Nevertheless in settings where these Assumptions are questionable, we

can still use the underlying moment inequalities below as a basis for partial identification and inference: for all j, t:

$$E[\delta_{jt}^u(s_t, \lambda_0) - x_{jt}'\beta_0 | z_{jt}] \geq 0$$
$$E[\delta_{jt}^\ell(s_t, \lambda_0) - x_{jt}'\beta_0 | z_{jt}] \leq 0.$$

Based on this conditional moment inequality model, One can use the method developed in Andrews and Shi (2013) to construct a joint confidence set for the full vector θ_0. This confidence set is constructed by inverting an Anderson-Rubin test: $CS = \{\theta : T(\theta) \leq c(\theta)\}$ for some test statistic $T(\theta)$ and critical value $c(\theta)$. Computing this set amounts to computing the 0-level set of the function $T(\theta) - c(\theta)$, where $c(\theta)$ typically is simulated quantiles and thus a non-smooth function of θ. A new approach that is computationally less burdensome when β is high dimensional is proposed in Gandhi, Lu, and Shi (2013), which also includes Monte Carlo simulations and empirical results using the profiling approach under partial identification.

6 Inference

In this section we discuss statistical inference based on our point estimator. We show that the estimator is asymptotically normal, which is a similar result to that in Kahn and Tamer (2009) for censored regression models.

More assumptions are needed. For clarity, we divide the assumptions into two groups, the first being standard ones similar to those in Freyberger (2015) and the second being the special assumptions that are needed to account for the presence and the unknown identity of the risky products. Let $B_c(\lambda_0)$ denote a open ball around λ_0 of radius $c > 0$.

Assumption 9. (a) *θ_0 is in the interior of Θ.*

(b) *The function $\delta_{jt}(\pi, \lambda)$ is twice-continuously differentiable in (π, λ) on $\Delta^1_{J_t} \times \Lambda$, for all j, t.*

(c) *For any sequence λ_T such that $\lambda_T - \lambda_0 \to_p 0$,*

$$T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \| \left[\delta_{jt}(\pi_t, \lambda_T) - \delta_{jt}(\pi_t, \lambda_0)\right] 1(z_{jt} \in \mathcal{Z}_0) \| = O_p(1) \|\lambda_T - \lambda_0\|,$$

$$T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \|\delta_{jt}(\pi_t, \lambda_T) - \delta_{jt}(\pi_t, \lambda_0)\| = O_p(\log(T)) \|\lambda_T - \lambda_0\|,$$

$$T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \|\delta_{jt}(\tilde{s}_t, \lambda_T) - \delta_{jt}(\tilde{s}_t, \lambda_0)\| = O_p(\log(T)) \|\lambda_T - \lambda_0\|.$$

(d) $\lim_{T\to\infty} \underline{n}_T^{-1} T^{1/2} = \lim_{T\to\infty} T^{-1} \overline{n}_T^{1/2} = 0$.

Let $\mathcal{G} \backslash \mathcal{G}_0$ denote the relative complement of $\mathcal{G}_0$ in $\mathcal{G}$. Let $\partial m_{jt}(\lambda)$ denote $\begin{pmatrix} \partial \delta_{jt}(\pi_t, \lambda)/\partial \lambda \\ x_{jt} \end{pmatrix}$.

Let $\Gamma_T(g) = T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t} E[\partial m_{jt}(\lambda_0)g(z_{jt})]$.

Assumption 10. (a) *There exists a constant $\eta > 0$ such that for all sufficiently small $c > 0$ and all T, we have*

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0 : T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t} E[(\log(s_{jt}+\iota_u/n_t)-\log(\pi_{jt}))g(z_{jt})]\le c} \mu(g) < c^\eta,$$

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0 : T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t} E[(\log(s_{jt}+\bar{\iota}_\ell/n_t)-\log(\pi_{jt}))g(z_{jt})]\ge -c} \mu(g) < c^\eta,$$

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0 : T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t} E[g(z_{jt})(n_t s_{jt}+\iota_u)^{-1}]\le c} \mu(g) < c^\eta.$$

(b) *When Assumption 5 holds, assume that*

$$\sup_{j,t}\sup_{\lambda:\|\lambda-\lambda_0\|\le c} E\left[\left\|\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda)}{\partial\pi}\right\|^2\right] < \infty$$

and

$$\sup_{j,t}\sup_{\lambda:\|\lambda-\lambda_0\|\le c}\sup_{\pi:\|\pi-\pi_t\|\le c}\left\|\frac{\partial^2\hat{\delta}_{jt}(\pi,\lambda)}{\partial\pi\partial\pi'}\right\| = O_p(1)$$

for some $c > 0$. When Assumption 6 holds, assume that

$$\sup_{j,t}\sup_{\lambda:\|\lambda-\lambda_0\|\le c} E\left[\left\|\frac{\partial\delta_{jt}(\pi_t,\lambda)}{\partial\pi}\right\|^2 1(z_{jt}\in\mathcal{Z}_0)\right] < \infty$$

and

$$\sup_{j,t}\sup_{\lambda:\|\lambda-\lambda_0\|\le c}\sup_{\pi:\|\pi-\pi_t\|\le c}\left\|\frac{\partial^2\delta_{jt}(\pi,\lambda)}{\partial\pi\partial\pi'}1(z_{jt}\in\mathcal{Z}_0)\right\| = O_p(1)$$

for some $c > 0$.

(c) $\sup_{j,t}\sup_{z_0\in\mathcal{Z}_0} E[\|\partial m_{jt}(\lambda_0)\|^2|z_{jt}=z_0] < \infty$ *and* $\sup_{j,t}\sup_{\lambda:\|\lambda-\lambda_0\|\le c}\left\|\frac{\partial\delta_{jt}(\pi_t,\lambda)}{\partial\pi\partial\pi'}1(z_{jt}\in\mathcal{Z}_0)\right\| = O_p(1)$ *for some $c > 0$.*

(d) $\lim_{T\to\infty}\sum_{g\in\mathcal{G}_0}\mu(g)\Gamma_T(g)\Gamma_T(g)' = \Upsilon$ *for a matrix Υ of full rank.*

(e) $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\sum_{g,g^*\in\mathcal{G}_0} Cov\left(\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt}), \sum_{j=1}^{J_t}\xi_{jt}g^*(z_{jt})\right)\Gamma_T(g)\Gamma_T(g)'\mu(g)\mu(g^*) = V$.

Theorem 3. *Suppose that Assumptions 1, 4, and 7-10 hold. Also suppose that either Assumption 5 or Assumption 6 holds. Then we have*

$$\sqrt{T}(\widehat{\theta}_T - \theta_0) \to_d N(0, \Upsilon^{-1}V\Upsilon^{-1}).$$

Remark 1. Note that Υ and V depend on $\mathcal{G}_0$ which in turn depends on the unknown set $\mathcal{Z}_0$. Thus, estimating the asymptotic variance covariance matrix can be difficult. Instead, following Kahn and Tamer (2009), we recommend using non-parametric bootstrap to obtain standard errors and confidence intervals. We follow this recommendation in the empirical application in Section 8. We also evaluate the performance of bootstrap standard errors and several bootstrap-based confidence intervals in our Monte Carlo experiments in Section 7.

Remark 2. The asymptotic variance formula also makes it clear that the choice of instrumental function set $\mathcal{G}$ affects estimation accuracy. Potentially, one could choose $\mathcal{G}$ to minimize the asymptotic variance. However, the theory for which does not seem to resemble the existing efficiency theory for conditional moment equalities, e.g. Chamberlain (1987), Newey (1990), and Ai and Chen (2003), mainly due to the structure that $\mathcal{G}$ needs to take to preserve the information in the conditional moment inequalities. We thus leave this for future research.

7 Monte Carlo Simulations

In this section, we present two sets of Monte Carlo experiments with random coefficient logit models. The first experiment investigates the performance of our approach with moderate fractions of zero shares, which should cover most of the empirical scenarios. In the second experiment, we test our estimator with a data generating process that produces extremely large fractions of zeros; the purpose is to further illustrate the key idea of our estimator in exploiting the long tail pattern that is naturally present in the data.

Both experiments use the a random coefficient logit model, where the utility of consumer i for product j in market t is

$$u_{ijt} = \alpha_0 + x_{jt}\beta_0 + \lambda_0 x_{jt} v_i + \xi_{jt} + \epsilon_{ijt},$$

where $v_i \sim N(0,1)$, λ_0 is the standard deviation of the random coefficients on x_{jt}, ϵ_{ijt}'s are i.i.d. across i, j and t following Type I extreme value distribution. The parameters of interest are β_0 and λ_0, while α_0 is a nuisance parameter. In both experiments, we set $\lambda_0 = .5$, $\beta_0 = 1$ and vary α_0 for different designs. We simulate T markets, each with J products.

7.1 Moderately Many Zeroes

In the first experiment, the observed and unobserved characteristics are generated as $x_{jt} = \frac{j}{10} + N(0,1)$ and $\xi_{jt} \sim N(0,.1^2)$ for each product j in market t. Thus one feature of the design is that the x_{jt} has some persistence across markets - products with larger index tend to have higher value of x (which respects the nature of the variation in the scanner data shown in Section 2). Finally, the vector of empirical shares in market t, $(s_{0t}, s_{1t}, ..., s_{Jt})$, is generated from Multinomial $\left(n, [\pi_{0t}, \pi_{1t}, ..., \pi_{Jt}]'\right) \Big/ n$, where n represents the number of consumers in each mar-

ket.[13]

With the simulated data set $\{(s_{jt}, x_{jt}) : j = 1, ..., J\}_{t=1}^{T}$, we compute our bound estimator (bound), the standard BLP estimator using s_t in place of π_t and discarding observations with $s_{jt} = 0$ (ES), the standard BLP estimator using $\tilde{s}_t$ (no zeros) in place of π_t (LS).

All the estimators require simulating the market shares and solving demand systems for each trial of λ in optimizing the objective function for estimation. We use the same set of random draws of v_i as in the data generating process to eliminate simulation error as it is not the focus of this paper. BLP contraction mapping method is employed to numerically solve the demand systems.

We simulate 1000 datasets $\{(s_t^r, x_t^r) : t = 1, ..., T\}_{r=1}^{1000}$ and implement all the estimators mentioned above on each for a repeated simulation study. For the instrumental functions, we use the countable hyper-cubes defined in (4.9), and set $\bar{r}_T = 50$. The choices of ι_ℓ and ι_u follow Subsection 4.2. For the BLP estimator, we use $\left(1, x_{jt}, x_{jt}^2 - 1, x_{jt}^3 - 3x_{jt}\right)$ (the first three Hermite polynomials) as instruments to construct the GMM objective function. Alternative transformations of x_{jt} as instruments yield effectively the same results.

The bias and standard deviation of the estimators are presented in *Table* 2. As we can see from the table, The standard estimator with s_t shows large bias for both β and λ. Replacing the empirical share s_t with the Laplace share $\tilde{s}_t$ (and thus not discarding the observations with $s_{jt} = 0$) increases the bias for β although reducing the bias for λ. Our bound estimators are the least biased, and its bias is very small for both parameters, especially when the sample size (T) is larger.

[13] The π_t has no closed form solution in the random coefficient model, and thus, we compute them via simulation, i.e.,

$$\pi_{jt} = \frac{1}{s} \sum_{i=1}^{s} \frac{\exp\left(\alpha_0 + x_{jt}\beta_0 + \lambda_0 x_{jt} v_i + \xi_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(\alpha_0 + x_{kt}\beta_0 + \lambda_0 x_{kt} v_i + \xi_{kt}\right)},$$

where $s = 1000$ is the number of consumer type draws (v_i).

Table 2: Monte Carlo Results: Estimation

DGP	T	Ave. % of Zeros		ES		LS		Bound	
				λ	β	λ	β	λ	β
I	25	9.52%	Bias	.3718	-.1941	.2900	-.2167	.0432	-.0441
			SD	.0337	.0160	.0221	.0115	.0475	.0351
	50	9.48%	Bias	.3712	-.1939	.2912	-.2172	.0203	-.0242
			SD	.0236	.0118	.0164	.0082	.0397	.0293
	100	9.46%	Bias	.3714	-.1941	.2900	-.2169	.0027	-.0087
			SD	.0169	.0081	.0112	.0055	.0314	.0237
II	25	18.55%	Bias	.6752	-.6115	.4023	-.4675	.0168	-.0326
			SD	.0845	.0655	.0315	.0229	.0534	.0540
	50	18.54%	Bias	.6649	-.6040	.3993	-.4657	-.0053	-.0056
			SD	.0580	.0462	.0223	.0158	.0412	.0415
	100	18.50%	Bias	.6624	-.6021	.3983	-.4651	-.0123	.0042
			SD	.0422	.0333	.0163	.0114	.0299	.0300
III	25	41.14%	Bias	.7302	-1.3220	.3868	-.9863	-.0325	.0225
			SD	.2022	.2890	.0366	.0460	.0483	.0722
	50	41.11%	Bias	.7092	-1.2947	.3830	-.9819	-.0291	.0252
			SD	.1373	.1975	.0262	.0323	.0373	.0549
	100	41.10%	Bias	.7070	-1.2935	.3809	-.9794	-.0178	.0123
			SD	.0911	.1325	.0188	.0232	.0283	.0392
IV	25	52.38%	Bias	.4013	-1.1035	.2907	-1.1412	-.0451	.0440
			SD	.1346	.2435	.0304	.0453	.0536	.0916
	50	52.35%	Bias	.3942	-1.0937	.2877	-1.1369	-.0300	.0262
			SD	.0956	.1740	.0214	.0313	.0403	.0652
	100	52.36%	Bias	.3916	-1.0901	.2862	-1.1349	-.0168	.0094
			SD	.0687	.1255	.0154	.0227	.0313	.0478

Note: 1. $J = 50$, $N = 10,000$, $\beta_0 = 1$, $\lambda_0 = .5$, Number of Repetitions = 1000.

2. "ES": Empirical Shares; "LS": Laplace Shares.

3. DGP: I, II, III and IV correspond to$\alpha_0 = -9$, -10, -12 and -13, respectively.

Next, we examine the performance of our proposed bootstrap procedure and the results are reported in Table 3. We can see that bootstrap standard errors are on average slightly larger than the standard deviation of the estimators, especially for the cases with large fraction of zeros and small sample size. Also, we compute two versions of bootstrap confidence intervals and find that the "Normal CI", based on normal quantile and bootstrap standard errors, outperforms the standard nonparametric bootstrap confidence interval and gets very close to the nominal level (95%) of coverage probability as the sample size gets large.

Table 3: Monte Carlo Results: Bootstrap

DGP	T	Ave. % of Zeros	Actual SD		BS SE		CP: BS CI		CP: Normal CI	
			λ	β	λ	β	λ	β	λ	β
I	25	9.52%	.0473	.0350	.0471	.0351	.8408	.8178	.8579	.7608
	50	9.48%	.0395	.0291	.0399	.0299	.8560	.8560	.9340	.8880
	100	9.46%	.0312	.0235	.0324	.0244	.8418	.8571	.9592	.9459
II	25	18.54%	.0531	.0537	.0564	.0586	.8490	.8730	.9670	.9460
	50	18.54%	.0411	.0415	.0425	.0434	.8170	.8480	.9560	.9680
	100	18.49%	.0299	.0300	.0312	.0314	.8629	.8873	.9350	.9645
III	25	41.13%	.0487	.0730	.0545	.0851	.7920	.8450	.9270	.9740
	50	41.09%	.0378	.0554	.0393	.0586	.8550	.8950	.8970	.9430
	100	41.09%	.0285	.0394	.0293	.0420	.8846	.9231	.9180	.9585
IV	25	52.39%	.0539	.0913	.0560	.0988	.8120	.8730	.8870	.9520
	50	52.35%	.0402	.0644	.0427	.0718	.8696	.9188	.9147	.9639
	100	52.36%	.0317	.0483	.0317	.0506	.8444	.9010	.9242	.9465

Note: 1. All the settings are identical to Table 1.

2. "BS SE" refers to average bootstrap standard error.

3. "CP: BS CI" refers to the coverage probability of the 95% nonparametric bootstrap CI.

4. "CP: Normal CI" refers to the coverage probability of the 95% normal CI with bootstrap s.e.

7.2 Extremely Many Zeroes

Next we pressure test our bound estimator by pushing the fraction of zeroes in empirical shares toward the extreme. We modify the DGP slightly to produce very high fraction of zeros. Specifically, we generate x_{jt} from the following discrete distribution

x	1	12	15
$\Pr(x_{jt} = x)$	.99	.005	.005

and

$$\xi_{jt} \sim 1(x_{jt} = 1) \times N\left(0, 2^2\right) + 1(x_{jt} \neq 1) \times N\left(0, .1^2\right).$$

All the other aspects of the DGP is the identical to the previous DGP.

The fractions of zeroes are made very high: 82%-96% by choosing the α_0 parameter. With such high fractions of zeroes, the vast majority of observations are uninformative. Thus, we need larger sample size for any estimator to perform well. We consider $T = 100, 200, 400$. For simplicity of presentation and to reduce computational burden, we will here fix λ at its true value, and only investigate the behavior of the estimators for β .

The results are reported in *Table* 4, and they are very encouraging for the bound approach. The ES estimator is severely biased toward 0, so is the LS estimator. The bound estimator is remarkably accurate in these extreme cases. The performance highlights the key idea of identification behind our estimator: utilizing the information in safe products with inherently thick demand to identify the model while controlling the risky products with small/zero sales properly.

Table 4: Monte Carlo Results: Very Large Fraction of Zeros

DGP	T	Ave. % of Zeros		ES	LS	Bound
I	100	84.73%	Bias	-.2698	-.2643	-.0076
			SD	.0060	.0058	.0123
	200	84.68%	Bias	-.2695	-.2640	-.0073
			SD	.0042	.0040	.0093
	400	84.71%	Bias	-.2692	-.2639	-.0066
			SD	.0030	.0030	.0066
II	100	91.45%	Bias	-.3328	-.3319	-.0072
			SD	.0066	.0061	.0121
	200	91.40%	Bias	-.3327	-.3317	-.0072
			SD	.0047	.0043	.0091
	400	91.41%	Bias	-.3319	-.3314	-.0058
			SD	.0036	.0033	.0066
III	100	95.37%	Bias	-.3992	-.4028	-.0065
			SD	.0079	.0070	.0126
	200	96.36%	Bias	-.3991	-.4025	-.0065
			SD	.0056	.0049	.0093
	400	96.35%	Bias	-.3986	-.4023	-.0061
			SD	.0040	.0035	.0065

Note: 1. $T = 100$, $J = 50$, $N = 10,000$, $\beta_0 = 1$, $\lambda_0 = .5$, Number of Repetitions = 1000.

2. We fix $\lambda = \lambda_0$ (at the true value) without estimating it.

3. DGP: I, II, III correspond to $\alpha_0 = -13, -14, -15$.

8 Empirical Application

In this section, we apply our estimator on the DFF scanner data previewed in Section 2. In particular, we focus on the canned tuna category, as previously studied by Chevalier, Kashyap, and Rossi (2003) (CKR for short) and Nevo and Hatzitaskos (2006) (NH for short). CKR observed using the same data discussed in Section 2 that the share weighted price of tuna fell by 15 percent during Lent, which is a high demand period for this product. They attributed the outcome to loss-leading behavior on the part of retailers. NH on the other hand suggest that this pricing pattern in the tuna data could instead be explained by increased price sensitivity of consumers (consistent with an increase in search) which causes a re-allocation of market shares towards less expensive products in the Lent period, and hence a fall in the observed share weighted price index. They test this hypothesis directly in the data by estimating demand parameters separately in the Lent and Non-Lent periods, and find that demand becomes more elastic in the high demand (Lent) period.

Here we revisit the groundwork laid by NH to examine the difference in price elasticity between Lent and non-Lent periods. The main difference in our analysis is that we use data on all products in the analysis, while NH restrict the sample to include only the top 30 UPCs and thus automatically drop products with small/zero sales. There are two main questions we seek to address: (a) Does

the selection of UPC's with only positive shares significantly bias the estimates of price elasticity and (b) Does the difference in price elasticities between the Lent and Non-Lent period persist after properly controlling for zeroes.

To make the comparison clear, we use largely the same specification of the model used in NH. In particular we consider a logit specification

$$u_{ijt} = \alpha p_{jt} + \beta x_{jt} + \xi_{jt} + \epsilon_{ijt},$$

where the control variables x_{jt} consist of UPC fixed effects and a time trend.[14] The week to week variation in the product-/market-level unobserved demand shock ξ_{jt} largely captures the short-term promotional efforts, e.g., in-store advertising and shelving choices, because the UPC fixed effects control the intrinsic product quality that is likely to be stable over short time horizon. Since stores are likely to advertise or shelf the product in a more prominent way during weeks when the product is on a price sale, we expect a negative correlation between price and the unobservable. We construct instruments for price by inverting DFF's data on gross margin to calculate the chain's wholesale costs, which is the standard price instrument in the literature that has studied the DFF data.[15]

We implement our bound estimator defined by (4.7) to obtain point estimate of (α, β) in the model. And the 95% confidence interval for the parameters are obtained using nonparametric bootstrap.[16]

The estimation results are presented in Tables 5 and 6. [17] Table 5 shows that standard logit estimator that inverts empirical shares to recover mean utilities (and hence drops zeroes) has a significant selection bias towards zero. The UPC level elasticities for the logit model are small in economic magnitude, with the average elasticity in the data being -.572. Furthermore, over 90% of products having inelastic demand. Using our bounds approach instead to control for zeroes has a major effect on the estimated elasticities. Average demand elasticity for UPC's becomes -1.51 and less than 30% percent of observations have inelastic demand. This change in the direction of elasticities is consistent with the attenuation bias effects of dropping products with small/zero

[14] Empirical market shares are constructed using quantity sales and the number of people who visited the store that week (the customer count) as the relevant market size.

[15] The gross margin is defined as (retail price - wholesale cost)/retail price, so we get wholesale cost using retail price×(1 - gross margin). The instrument defensible in the store disaggregated context we consider here because it has been shown that price sales in retail price primarily reflect a reduction in retailer margins rather than a reduction in marginal costs (see e.g., Chevalier, Kashyap, and Rossi (2003) and Hosken and Reiffen (2004)). Thus sales (and hence promotions) are not being driven by the manufacturer through temporary reduction in marginal costs.

[16] The procedure contains the following steps: (1) draw with replacement a bootstrap sample of *markets*, denoted as $\{t_1, ..., t_T\}$; (2) compute the bound estimator $\widehat{\theta}_T^{BD*}$ using the bootstrap sample; (3) repeat (1)-(2) for B_T times and obtain B_T independent (conditional on the original sample) copies of $\widehat{\theta}_T^{BD*}$; (4) $q_T^*(\tau)$ is the τ-th quantile of the B_T copies of $\left(\widehat{\theta}_T^{BD*} - \widehat{\theta}_T^{BD}\right)$, then the 95% bootstrap confidence interval is $\left[\widehat{\theta}_T^{BD} - q_T^*(.975), \widehat{\theta}_T^{BD} - q_T^*(.025)\right]$.

[17] In principle we can estimate our model separately for each store, letting preferences change freely over stores depending on local preferences. These results are available upon request. Here we present for the results of demand pooling together all stores together as was done by Nevo and Hatzitaskos (2006). The store level regressions results are very similar to the pooled store regression and the latter is a more concise summary of demand behavior that we present here.

market shares.

Table 5: Demand Estimation Results

	BLP	Bound
Price Coefficient	-.39	-1.03
95% CI	[-.40, -.38]	[-1.92, -.91]
Ave. Own Price Elasticity	-.57	-1.51
Fraction of Inelastic Products	90.04%	28.20%
No. of Obs.	862,683	959,331

Table 6: Demand in Lent vs. Non-Lent

	BLP		Bound	
	Lent	Non-Lent	Lent	Non-Lent
Price Coefficient	-.518	-.371	-1.23	-.75
95% CI	[-55, -.48]	[-.38, -.36]	[-1.70, -.92]	[-1.12, -.33]
Ave. Own Price Elasticity	-.757	-.544	-1.80	-1.10
Fraction of Inelastic Products	84.02%	92.84%	16.79%	43.94%
No. of Obs.	70,496	792,187	78,838	880,493

Our second result is that demand becomes more elastic in the high demand period, as shown in Table 6. This is consistent with Nevo and Hatzitaskos (2006)'s findings that are based on the standard logit estimator with zeroes being dropped. However, the Lent effect is bigger according to our bounds estimator that controls for the zeroes. In other words, correcting the selection bias, our bound estimator brings the price coefficient and elasticity higher and the correction effect is higher for the Lent period than for the non-Lent period. Since the fractions of zeroes are remarkably close between Lent and non-Lent periods, we suspect that the difference in the correction effect is due to a change in the distribution of the unobservable ξ.

To further investigate this, we first replicate the reduced form finding of Nevo and Hatzitaskos (2006) that suggested a change in price sensitivity in the Lent period. This is reported in Table 7, which shows that although the price index of tuna during Lent appears to be approximately 15 percent less expensive than other weeks (as previously underscored by CKR), the average price of tuna is virtually unchanged between the Lent versus non-Lent period. Hence it is a re-allocation of demand towards less expensive products during Lent that drives the change in the aggregate price index.

Table 7: Regression of Price Index on Lent

	P	$\bar{P}$
	(Price Index)	(Average Price)
Lent	-.150	-.009
s.e.	(.0005)	(.0003)

We take this decomposition one step further than NH, and examine the price index separately

for products "on sale" and "regularly priced" during these periods.[18] As can be seen in Table 8, it is the sales price index that is the key driver of the aggregate price index being cheaper during Lent. However the average price of an "on-sale" product is not cheaper in the Lent period. This shows that it is a re-allocation towards more steeply discounted "on-sale" product during Lent that is driving this change in the aggregate price index. But we do not see a corresponding such reallocation for "regularly priced" products.

Table 8: Regression of Sales Price Index on Lent

	P (Price Index)		$\bar{P}$ (Average Price)	
	Sale	Regular	Sale	Regular
Lent	-.199	.035	.010	.001
s.e.	(.0017)	(.0003)	(.0016)	(.0003)

This suggests a tighter coordination of promotional effort and discounting in the high demand period. In effect more steeply discounted products are receiving larger promotional effort on the part of the retailer during the high demand, which is similar in spirit to the loss-leader hypothesis originally advanced for this data by CKR. Since promotional effort in the model is largely captured through the unobservable ξ, this change in behavior of the unobservable would account for the selection effect due to dropping zeroes changing across the two periods: during Lent period, the variance of promotional effort is larger so the selection bias is worse. Hence, our results suggest that both demand and supply side effects contribute to the falling price during high demand period, which complement the findings of NH and CKR.

9 Conclusion

We have shown that differentiated product demand models have enough content to construct a system of moment inequalities that can be used to consistently estimate demand parameters despite a possibly large presence of observations with zero market shares in the data. We construct a GMM-type estimator based on these moment inequalities that is consistent and asymptotically normal under assumptions that are a reasonable approximation to the DGP in many product differentiated environments. Our application to scanner data reveals that taking the market zeroes in the data into account has economically important implications for price elasticities.

A key message from our analysis is that it is critical to not ignore the zero shares when estimating discrete choice models with disaggregated market data. And a potentially fruitful area for future research is the application of our approach to individual level choice data, such as a household panel. Aggregating over households is still necessary to control for price endogeneity, such as described by Berry, Levinsohn, and Pakes (2004) and Goolsbee and Petrin (2004), and thus zero market shares when we aggregate over limited sample of households in the data is a clear problem for

[18] We flag an observation in the data as being on sale if that particular UPC in that particular store in that particular week has at least a 5% reduction from highest price of previous 3 weeks.

many contexts. Nevertheless the demographic richness in the household panel provides additional identifying power for random coefficients. The approach we describe can offer a novel solution to the joint problem of endogenous prices and flexible consumer heterogeneity with micro data, which we plan to pursue in future work.

Appendix

A Further Illustrations of Zipf's Law

In Figure 4 we illustrate this regularity using data from the two other applications that were mentioned in Section 2: homicide rates and international trade flows. The left hand graph shows the annual murder rate (per 10,000 people) for each county in the US from 1977-1992 (for details about the data see Dezhbakhsh, Rubin, and Shepherd (2003)). The right hand side graph shows the import trade flows (measured in millions of US dollars) among 160 countries that have a regional trade agreement in the year 2006 (for details about the data see Head, Mayer, et al. (2013)). In each of these two cases we see the characteristic pattern of Zipf's law - a sharp decay in the frequency for large outcomes and a large mass near zero (with a mode at zero in each case).

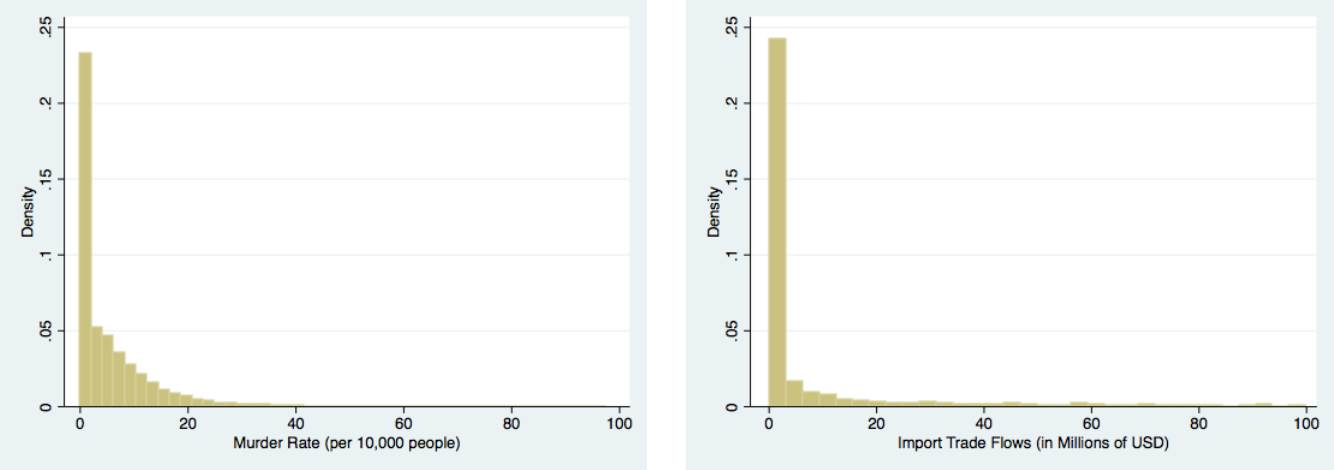


Figure 4: Zipf's Law in Crime and Trade Data

B Proofs of the Theorems

B.1 Proof of Theorem 1

Let

$$\widehat{Q}_{0,T}(\theta) = \sum_{g \in \mathcal{G}_0} \left\{ \left([\bar{m}_T^u(\theta, g)]_-^2 + \left[\bar{m}_T^\ell(\theta, g) \right]_+^2 \right) \mu(g) \right\}.$$

Proof of Theorem 1. First note that $\widehat{Q}_T(\theta_0) = \sum_{g \in \mathcal{G}} \mu(g) [\bar{m}_T^u(\theta_0, g)]_-^2 + \sum_{g \in \mathcal{G}} \mu(g) [\bar{m}_T^\ell(\theta_0, g)]_+^2$. Thus, Assumption 2(a) directly implies that

$$\widehat{Q}_T(\theta_0) = o_p(1). \tag{B.1}$$

Below we show that

$$\sup_{\theta\in\Theta}|\sqrt{\widehat{Q}_{0,T}(\theta)}-\sqrt{\widehat{Q}^*_T(\theta)}|=o_p(1). \tag{B.2}$$

Consider an arbitrary $c>0$. The theorem is implied by the following derivation:

$$\begin{aligned}
\Pr\left(\|\widehat{\theta}_T-\theta_0\|>c\right) &\le \Pr\left(\sqrt{\widehat{Q}^*_T(\widehat{\theta}_T)}\ge\sqrt{C(c)}\right)\\
&= \Pr\left(\sqrt{\widehat{Q}^*_T(\widehat{\theta}_T)}-\sqrt{\widehat{Q}_{0,T}(\widehat{\theta}_T)}+\sqrt{\widehat{Q}_{0,T}(\widehat{\theta}_T)}\ge\sqrt{C(c)}\right)\\
&\le \Pr\left(\sup_{\theta\in\Theta}|\sqrt{\widehat{Q}_{0,T}(\theta)}-\sqrt{\widehat{Q}^*_T(\theta)}|+\sqrt{\widehat{Q}_{0,T}(\widehat{\theta}_T)}\ge\sqrt{C(c)}\right)\\
&\le \Pr\left(\sup_{\theta\in\Theta}|\sqrt{\widehat{Q}_{0,T}(\theta)}-\sqrt{\widehat{Q}^*_T(\theta)}|+\sqrt{\widehat{Q}_{T}(\widehat{\theta}_T)}\ge\sqrt{C(c)}\right)\\
&\le \Pr\left(\sup_{\theta\in\Theta}|\sqrt{\widehat{Q}^0_T(\theta)}-\sqrt{\widehat{Q}^*_T(\theta)}|+\sqrt{\widehat{Q}_{T}(\theta_0)}\ge\sqrt{C(c)}\right)\\
&\le \Pr\left(\sup_{\theta\in\Theta}|\sqrt{\widehat{Q}^0_T(\theta)}-\sqrt{\widehat{Q}^*_T(\theta)}|\ge\sqrt{C(c)}/2\right)+\Pr\left(\widehat{Q}_T(\theta_0)\ge C(c)/4\right)\\
&\to 0,
\end{aligned} \tag{B.3}$$

where the first inequality holds by Assumption 3, the third inequality holds because $\widehat{Q}_T(\widehat{\theta}_T)$ differs from $\widehat{Q}_{0,T}(\widehat{\theta}_T)$ only in that the former takes the summation over a larger range, the fourth inequality holds because $\widehat{Q}_T(\widehat{\theta}_T)\le\widehat{Q}_T(\theta_0)$ by the definition of $\widehat{\theta}_T$ and the convergence holds by (B.1) and (B.2).

Now we show (B.2). Consider the derivation

$$\begin{aligned}
&\sup_{\theta\in\Theta}|\sqrt{\widehat{Q}_{0,T}(\theta)}-\sqrt{\widehat{Q}^*_T(\theta)}|\\
=&\sup_{\theta\in\Theta}\left|\sqrt{\sum_{g\in\mathcal{G}_0}\mu(g)\left\{[\bar{m}^u_T(\theta,g)]^2_-+[\bar{m}^\ell_T(\theta,g)]^2_+\right\}}-\sqrt{\sum_{g\in\mathcal{G}_0}\mu(g)\left\{\bar{m}_T(\theta,g)^2\right\}}\right|\\
\le&\sup_{\theta\in\Theta}\left|\sqrt{\sum_{g\in\mathcal{G}_0}\mu(g)\left\{\left(\sqrt{[\bar{m}^u_T(\theta,g)]^2_-+[\bar{m}^\ell_T(\theta,g)]^2_+}-|\bar{m}_T(\theta,g)|\right)^2\right\}}\right|\\
\le&\sup_{\theta\in\Theta}\left|\sqrt{\sum_{g\in\mathcal{G}_0}\mu(g)\left\{([\bar{m}^u_T(\theta,g)]_--[\bar{m}_T(\theta,g)]_-)^2+([\bar{m}^\ell_T(\theta,g)]_+-[\bar{m}_T(\theta,g)]_+)^2\right\}}\right|\\
\le&\sup_{\theta\in\Theta}\left|\sqrt{\sum_{g\in\mathcal{G}_0}\mu(g)\left\{(\bar{m}^u_T(\theta,g)-\bar{m}_T(\theta,g))^2+(\bar{m}^\ell_T(\theta,g)-\bar{m}_T(\theta,g))^2\right\}}\right|\\
\le&\sqrt{\sup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_0}|\bar{m}^u_T(\theta,g)-\bar{m}_T(\theta,g)|^2+\sup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_0}|\bar{m}^\ell_T(\theta,g)-\bar{m}_T(\theta,g)|^2}\\
\to_p&0,
\end{aligned} \tag{B.4}$$

where the first inequality holds by the triangular inequality for the norm

$$\|a(\cdot)\| := \sqrt{\sum_{g\in\mathcal{G}_0} \mu(g)a(g)^2 / \sum_{g\in\mathcal{G}_0} \mu(g)},$$

the second inequality holds by the triangular inequality for the Euclidean norm, the third inequality holds because $|[x]_- - [y]_-| \leq |x-y|$ and $[x]_+ = [-x]_-$, and the fourth inequality holds because $\mu : \mathcal{G} \to [0,1]$ is a probability measure on $\mathcal{G}$ and $\mathcal{G}_0 \subseteq \mathcal{G}$, and the convergence holds by Assumption 2(b). Therefore (B.2) is proved. □

B.2 Proof of Theorem 2

Proof of Theorem 2. First we show part (i). Let $\sup_{j,t:z_{jt}\in\mathcal{Z}_0}$ abbreviate $\sup_{t=1,\ldots,T}\sup_{j=1,\ldots,J_t:z_{jt}\in\mathcal{Z}_0}$. Consider the derivation:

$$\begin{aligned}
&\sup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_0} |\bar{m}_T^u(\theta,g) - \bar{m}_T(\theta,g)| \\
&= \sup_{\lambda\in\Lambda}\sup_{g\in\mathcal{G}_0} \left| \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\delta_{jt}^u(s_t,\lambda) - \delta_{jt}(\pi_t,\lambda))g(z_{jt}) \right| \\
&\leq \bar{J}\sup_{\lambda\in\Lambda}\sup_{j,t:z_{jt}\in\mathcal{Z}_0} |\delta_{jt}^u(s_t,\lambda) - \delta_{jt}(\pi_t,\lambda)| \\
&\leq \bar{J}\sup_{j,t:z_{jt}\in\mathcal{Z}_0} (|\log(s_{jt}+\iota_u/n_t) - \log(\tilde{s}_{jt})| + \bar{J}\sup_{\lambda\in\Lambda}\sup_{j,t:z_{jt}\in\mathcal{Z}_0} |\delta_{jt}(\tilde{s}_t,\lambda) - \delta_{jt}(\pi_t,\lambda)|,
\end{aligned}$$

where the first inequality holds by Assumption 7(d) and the definition of $\mathcal{G}_0$. Assumptions 4(a) and $0 < \iota_u < \infty$ together imply that $\sup_{j,t:z_{jt}\in\mathcal{Z}_0} |s_{jt} + \iota_u/n_t - \tilde{s}_{jt}| \to_p 0$. Moreover, Assumptions 4(a) and 7(g) together imply that

$$\sup_t \|\tilde{s}_t - \pi_t\| \to_p 0. \tag{B.5}$$

These and Assumptions 1(a) toghether imply that

$$\Pr\left(\inf_{j,t:z_{jt}\in\mathcal{Z}_0} \pi_{jt} > \underline{\varepsilon}_0, \inf_{j,t:z_{jt}\in\mathcal{Z}_0} s_{jt} + \iota_u/n_t > \underline{\varepsilon}_0/2, \inf_{j,t:z_{jt}\in\mathcal{Z}_0} \tilde{s}_{jt} > \underline{\varepsilon}_0/2 \right) \to 1.$$

This combined with Assumption 8 implies that $\sup_{\lambda\in\Lambda}\sup_{j,t:z_{jt}\in\mathcal{Z}_0} |\delta_{jt}(\tilde{s}_t,\lambda) - \delta_{jt}(\pi_t,\lambda)| \to_p 0$. Also, we have

$$\sup_{j,t:z_{jt}\in\mathcal{Z}_0} (|\log(s_{jt}+\iota_u/n_t) - \log(\tilde{s}_{jt})| \to_p 0.$$

because the logarithm function is uniformly continuous on the closed interval $[\underline{\varepsilon}_0/2, 1]$. Therefore, the first convergence in Assumption 2(b) holds. The second convergence holds by analogous arguments.

Now we show part (ii). We separate the two cases, one where Assumption 5 is satisfied and the other where Assumption 6 is satisfied.

Case 1: Assumption 5 is satisfied. In this case, the arguments for the first convergence and the second convergence in Assumption 2(a) are exactly analogous. Thus, we only discuss the first. Consider the derivation:

$$\begin{aligned}\bar{m}_T^u(\theta_0, g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}^u(s_t, \lambda_0) - x_{jt}'\beta_0)g(z_{jt}) \\ &\geq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt}) + \\ &\quad \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt} + \underline{\iota}_u/n_t) - \log(\pi_{jt}))g(z_{jt}) + \\ &\quad \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\hat{\delta}_{jt}(\tilde{s}_t, \lambda_0) - \hat{\delta}_{jt}(\pi_t, \lambda_0))g(z_{jt}), \end{aligned} \tag{B.6}$$

where the inequality holds because $\iota_u \geq \underline{\iota}_u$, $\xi_{jt} = \delta_{jt}(\pi_t, \lambda_0) - x_{jt}'\beta_0$ and $\hat{\delta}(\cdot, \lambda) = \delta(\cdot, \lambda) - \log(\cdot_j)$. We analyze the three summands one by one. For the first summand, observe that $\sum_{t=1}^{T}\sum_{j=1}^{J_t} E[\xi_{jt}^2] \leq \bar{J}MT$ by Assumption 7(f). We can then apply Lemma 4 in Appendix B.4 (with $w_{jt} = \xi_{jt}$) and get, for some constant C,

$$E\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\{\xi_{jt}g(z_{jt}) - E[\xi_{jt}g(z_{jt})]\}\right|^2 \leq \frac{C\bar{J}^2M}{T}. \tag{B.7}$$

The lemma applies due to Assumptions 7(d)-(e). Also, by Assumption 7(c), $E[\xi_{jt}g(z_{jt})] = 0$. Thus, we have

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt})\right| = O_p(T^{-1/2}). \tag{B.8}$$

Similar arguments apply to the second summand in (B.6) and yields

$$\begin{aligned} &E\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt} + \underline{\iota}_u/n_t) - \log(\pi_{jt}))g(z_{jt}) - E[(\log(s_{jt} + \underline{\iota}_u/n_t) - \log(\pi_{jt}))g(z_{jt})]\right|^2 \\ \leq &\frac{C\bar{J}^2}{T}\max_{j,t}E[(\log(s_{jt} + \underline{\iota}_u/n_t) - \log(\pi_{jt}))^2] \\ \leq &\frac{C\bar{J}^2}{T}\max_{t}[|\log(\underline{\iota}_u/n_t)|^2 + |\log(\underline{\varepsilon}_1/n_t)|^2] \\ \leq &\frac{2C\bar{J}^2(2(\log\overline{n}_T)^2 + (\log\underline{\iota}_u)^2 + (\log\underline{\varepsilon}_1)^2)}{T} \\ \to &0, \end{aligned} \tag{B.9}$$

where the second inequality holds by $s_{jt} \in [0,1]$ and Assumption 1(c) and the convergence holds by Assumptions 4(b) and 5(b). By the definition of $\underline{\iota}_u$, we have $E[(\log(s_{jt}+\underline{\iota}_u/n_t)-\log(\pi_{jt}))|\pi_{jt}, z_{jt}] \geq 0$, which then implies that $E[(\log(s_{jt}+\underline{\iota}_u/n_t)-\log(\pi_{jt}))g(z_{jt})] \geq 0$ for all $g \in \mathcal{G}$. Therefore, for any $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\log(s_{jt}+\underline{\iota}_u/n_t)-\log(\pi_{jt}))g(z_{jt}) < -c \right) = 0. \tag{B.10}$$

For the third summand in (B.6), consider the derivation

$$\begin{aligned} \sup_{g\in\mathcal{G}} \left| \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\hat{\delta}_{jt}(\tilde{s}_t,\lambda_0)-\hat{\delta}_{jt}(\pi_t,\lambda_0))g(z_{jt}) \right| &\leq \bar{J} \sup_{j,t} |\hat{\delta}_{jt}(\tilde{s}_t,\lambda_0)-\hat{\delta}_{jt}(\pi_t,\lambda_0)| \\ &\to_p 0, \end{aligned} \tag{B.11}$$

by (B.5) and Assumptions 5(a). Finally, (B.6), (B.8), (B.10), and (B.11) combined imply that for any $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \bar{m}_T^u(\theta_0, g) < -c \right) = 0,$$

which then implies the first convergence in Assumption 2(a) since $[\bar{m}_T^u(\theta_0,g)]_- = \max\{0, -\bar{m}_T^u(\theta_0,g)\}$.

Case 2: Assumption 6 is satisfied. We begin with the first convergence in Assumption 2(a). Consider the decomposition:

$$\begin{aligned} \bar{m}_T^u(\theta_0,g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\delta_{jt}^u(s_t,\lambda_0) - x_{jt}'\beta_0)g(z_{jt}) \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} \xi_{jt} g(z_{jt}) + \\ &\quad \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\log(s_{jt}+\iota_u/n_t)-\log(\tilde{s}_{jt}))g(z_{jt}) + \\ &\quad \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\delta_{jt}(\tilde{s}_t,\lambda_0)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt}). \end{aligned} \tag{B.12}$$

The first summand is $O_p(T^{-1/2})$ by (B.8). The second summand is nonnegative almost surely because $\tilde{s}_{jt} = s_{jt} + 1/n_t$ and $\iota_u \geq 1$ (Assumption 6(d)). For the third summand, similar to (B.9), we get for some generic constant C,

$$\begin{aligned} &E \sup_{g\in\mathcal{G}} \left| \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\delta_{jt}(\tilde{s}_t,\lambda_0)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt}) - E[(\delta_{jt}(\tilde{s}_t,\lambda_0)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt})] \right|^2 \\ \leq &\frac{C\bar{J}^2}{T} \max_{j,t} E[(\delta_{jt}(\tilde{s}_t,\lambda_0)-\delta_{jt}(\pi_t,\lambda_0))^2] \end{aligned}$$

$$\leq \frac{2CC_0\bar{J}^2\log(\overline{n}_T)^2}{T}$$
$$\to 0, \tag{B.13}$$

where the second inequality holds by Assumption 6(c) also using Assumptions 1(c) and 6(d), and the convergence holds by Assumption 4(b). Moreover, Assumption 6(b) implies that $E[(\delta_{jt}(\tilde{s}_t,\lambda_0) - \delta_{jt}(\pi_t,\lambda_0))g(z_{jt})] \geq 0$. This combined with (B.13) implies that, for any $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t,\lambda_0) - \delta_{jt}(\pi_t,\lambda_0))g(z_{jt}) < -c\right) = 0. \tag{B.14}$$

This combined with the arguments for the first two summands of (B.6) above yields: for any $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \bar{m}^u_T(\theta_0, g) < -c\right) = 0,$$

which then implies the first convergence in the statement of Assumption 2(a) because $[\bar{m}^u_T(\theta_0,g)]_- = \max\{0, -\bar{m}^u_T(\theta_0,g)\}$.

Now we show the second convergence in the statement of Assumption 2(a) for Case 2. Note that

$$\begin{aligned}
\bar{m}^\ell_T(\theta_0, g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta^\ell_{jt}(s_t,\lambda_0) - x'_{jt}\beta_0)g(z_{jt}) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt}) + \\
&\quad \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt} + \bar{\iota}_\ell/n_t) - \log(\pi_{jt}))g(z_{jt}) + \\
&\quad \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\hat{\delta}_{jt}(\tilde{s}_t,\lambda_0) - \hat{\delta}_{jt}(\pi_t,\lambda_0))g(z_{jt}),
\end{aligned} \tag{B.15}$$

where the inequality holds because $\iota_\ell \leq \bar{\iota}_\ell$ by Assumption 6(d). The first summand is $O_p(T^{-1/2})$ by (B.8). Since $\bar{\iota}_\ell > 0$, the arguments for (B.10) directly apply to the second summand to yield that, for any $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt} + \bar{\iota}_\ell/n_t) - \log(\pi_{jt}))g(z_{jt}) > c\right) = 0. \tag{B.16}$$

For the third summand in (B.15), we can apply the same arguments as those for (B.14) where we

use Assumption 6(a) in place of Assumption 6(b). Such arguments yield, for all $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} (\hat{\delta}_{jt}(\tilde{s}_t, \lambda_0) - \hat{\delta}_{jt}(\pi_t, \lambda_0)) g(z_{jt}) > c \right) = 0. \tag{B.17}$$

Therefore, for any $c > 0$,

$$\lim_{T\to\infty} \Pr\left(\inf_{g\in\mathcal{G}} \bar{m}_T^{\ell}(\theta_0, g) > c \right) = 0,$$

which then implies the second convergence in the statement of Assumption 2(a) because $[\bar{m}_T^{\ell}(\theta_0, g)]_+ = \max\{0, \bar{m}_T^{\ell}(\theta_0, g)\}$. □

B.3 Proof of Asymptotic Normality

To prove Theorem 3, we first give an auxiliary theorem that shows the convergence rate of $\widehat{\theta}_T$.

Theorem 4. *Suppose that Assumptions 1, 4, 7-10 hold. Also suppose that either Assumption 5 or Assumption 6 hold. Then we have* $\widehat{\theta}_T - \theta_0 = O_p(T^{-1/2})$.

Theorem 4 is proved using the following three lemmas. Theorem 4 and one of the lemmas together imply Theorem 3 as we explain immediately below. We give the proofs of Theorem 4 and the three lemmas in turn following the proof of Theorem 3.

Lemma 1. *Suppose that Assumptions 1, 4, 7-10 hold. Also suppose that either Assumption 5 or Assumption 6 hold. Then we have for any sequence* θ_T *such that* $\theta_T - \theta_0 = O_p(T^{-1/2})$, $\widehat{Q}_T(\theta_T) - \widehat{Q}_{0,T}(\theta_T) = o_p(T^{-1})$.

Lemma 2. *Suppose that Assumptions 1, 4, 7-10 hold and* $\iota_\ell, \iota_u \in [0, \infty)$. *Then we have*

(a) *for an open ball* $B_c(\theta_0)$ *of radius* $c > 0$ *around* θ_0, *we have* $\sup_{\theta\in B_c(\theta_0)} \left| \sqrt{\widehat{Q}_{0,T}(\theta)} - \sqrt{\widehat{Q}_T^*(\theta)} \right| = o_p(T^{-1/2})$, *and*

(b) $\widehat{Q}_T^*(\theta_0) = O_p(T^{-1})$.

Lemma 3. *Suppose that Assumptions 1, 4, 7-10 hold. For any sequence of random vectors* θ_T *such that* $\|\theta_T - \theta_0\| \to_p 0$, *we have*

(a) $\widehat{Q}_T^*(\theta_T) - \widehat{Q}_T^*(\theta_0) = (\theta_T - \theta_0)'\widehat{\Upsilon}_T(\theta_T - \theta_0) + 2W_T'(\theta_T - \theta_0) + o_p(1)\|\theta_T - \theta_0\|^2$, *where*

$$\widehat{\Upsilon}_T = \sum_{g\in\mathcal{G}_0} \mu(g)\widehat{\Gamma}_T(g)\widehat{\Gamma}_T(g)'$$

$$W_T = \sum_{g\in\mathcal{G}_0} \mu(g)\bar{m}_T(\theta_0, g)\widehat{\Gamma}_T(g)$$

$$\widehat{\Gamma}_T(g) = T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t} g(z_{jt})\partial m_{jt}(\lambda_0),$$

and

(b) $\widehat{\Upsilon}_T \to_p \Upsilon$ *and* $T^{1/2}W_T \to_d N(0, V)$.

Proof of Theorem 3. We use Theorem 2 of Sherman (1993) to prove the theorem. By Theorem 2 of Sherman (1993), the conclusion of our Theorem 3 holds under two conditions:

(i) $\|\widehat{\theta}_T - \theta_0\| = O_p(T^{-1/2})$,

(ii) uniformly over $O_p(T^{-1/2})$ neighborhood of θ_0, $\widehat{Q}_T(\theta) - \widehat{Q}_T(\theta_0) = (\theta - \theta_0)'\Upsilon(\theta - \theta_0) + 2T^{-1/2}B_T'(\theta - \theta_0) + o_p(T^{-1})$ for a random vector B_n such that $B_n \to_d N(0, V)$.

Condition (i) is implied by Theorem 4. To establish condition (ii), consider the derivation: for any sequence θ_T such that $\theta_T - \theta_0 = O_p(T^{-1/2})$,

$$\begin{aligned}
\widehat{Q}_T(\theta) - \widehat{Q}_T(\theta_0) &= [\widehat{Q}_T(\theta_T) - \widehat{Q}_{0,T}(\theta_T)] + [\widehat{Q}_{0,T}(\theta_T) - \widehat{Q}^*_T(\theta_T)] + \\
&\quad [\widehat{Q}^*_T(\theta_T) - \widehat{Q}^*_T(\theta_0)] + [\widehat{Q}^*_T(\theta_0) - \widehat{Q}_{0,T}(\theta_0)] + [\widehat{Q}_{0,T}(\theta_0) - \widehat{Q}_T(\theta_0)] \\
&= o_p(T^{-1}) + [\widehat{Q}_{0,T}(\theta_T) - \widehat{Q}^*_T(\theta_T)] + \\
&\quad [\widehat{Q}^*_T(\theta_T) - \widehat{Q}^*_T(\theta_0)] + [\widehat{Q}^*_T(\theta_0) - \widehat{Q}_{0,T}(\theta_0)] + o_p(T^{-1}),
\end{aligned} \tag{B.18}$$

where the second equality holds by Lemma 1. For the summand $[\widehat{Q}_{0,T}(\theta_T) - \widehat{Q}^*_T(\theta_T)]$, consider the derivation:

$$\begin{aligned}
\widehat{Q}_{0,T}(\theta_T) - \widehat{Q}^*_T(\theta_T) &= \left(\sqrt{\widehat{Q}_{0,T}(\theta_T)} - \sqrt{\widehat{Q}^*_T(\theta_T)}\right)^2 + 2\left(\sqrt{\widehat{Q}_{0,T}(\theta_T)} - \sqrt{\widehat{Q}^*_T(\theta_T)}\right)\left(\sqrt{\widehat{Q}^*_T(\theta_T)}\right) \\
&= o_p(T^{-1}) + o_p(T^{-1/2})\sqrt{\widehat{Q}^*_T(\theta_T) - \widehat{Q}^*_T(\theta_0) + \widehat{Q}^*_T(\theta_0)} \\
&= o_p(T^{-1}) + o_p(T^{-1/2})\sqrt{\widehat{Q}^*_T(\theta_T) - \widehat{Q}^*_T(\theta_0) + O_p(T^{-1})} \\
&= o_p(T^{-1}) + o_p(T^{-1/2})\sqrt{O_p(T^{-1}) + O_p(T^{-1})} \\
&= o_p(T^{-1}),
\end{aligned} \tag{B.19}$$

where the second equality holds by Lemma 2(a), the third equality holds by Lemma 2(b), and the fourth equality holds by Lemma 3(a)-(b). Similar arguments show that the summand $[\widehat{Q}_{0,T}(\theta_0) - \widehat{Q}^*_T(\theta_0)] = o_p(T^{-1})$. Therefore,

$$\widehat{Q}_T(\theta) - \widehat{Q}_T(\theta_0) = o_p(T^{-1}) + \widehat{Q}^*_T(\theta_T) - \widehat{Q}^*_T(\theta_0) \tag{B.20}$$

This combined with Lemma 3(a)-(b) shows the condition (ii) where $B_T = T^{1/2}W_T$. This concludes the proof of Theorem 3. □

Proof of Theorem 4. We prove Theorem 4 using Lemmas 1-3. The three lemmas imply that

$$\begin{aligned}
&(eig_{min}(\Upsilon) + o_p(1))\|\widehat{\theta}_T - \theta_0\|^2 + O_p(T^{-1/2})\|\widehat{\theta}_T - \theta_0\| \\
&\leq \widehat{Q}^*_T(\widehat{\theta}_T) - \widehat{Q}^*_T(\theta_0)
\end{aligned}$$

$$
\begin{aligned}
&\leq (\sqrt{\widehat{Q}_{0,T}(\widehat{\theta}_T)} + o_p(T^{-1/2}))^2 - \widehat{Q}^*_T(\theta_0) \\
&\leq 2\widehat{Q}_{0,T}(\widehat{\theta}_T) + o_p(T^{-1}) - \widehat{Q}^*_T(\theta_0) \\
&\leq 2\widehat{Q}_T(\widehat{\theta}_T) + o_p(T^{-1}) - \widehat{Q}^*_T(\theta_0) \\
&\leq 2\widehat{Q}_T(\theta_0) + o_p(T^{-1}) - \widehat{Q}^*_T(\theta_0) \\
&\leq 2(\widehat{Q}_T(\theta_0) - \widehat{Q}_{0,T}(\theta_0)) + 2\widehat{Q}_{0,T}(\theta_0) + o_p(T^{-1}) \\
&= O_p(T^{-1}),
\end{aligned} \tag{B.21}
$$

where $eig_{\min}(\Upsilon)$ is the smallest eigenvalue of Υ, the first equality holds by Lemma 3(a)-(b), the second inequality holds by Lemma 2(a), the third inequality holds by the algebraic inequality $(a+b)^2 \leq 2a^2+2b^2$, the fourth inequality holds because $\widehat{Q}_{0,T}(\cdot)$ and $\widehat{Q}_T(\cdot)$ are defined to be exactly the same, both being weighted sums of nonnegative terms, exept that the former sums over fewer terms, the fifth inequaity holds because $\widehat{\theta}_T$ is the minimizer of $\widehat{Q}_T(\cdot)$, the sixth inequality holds because $\widehat{Q}^*_T(\theta_0) \geq 0$, and the equality holds by Lemmas 1 and 2(a)-(b). Let ζ be an arbitrary positive number, we next show that we can find a constant M_1 large enough so that

$$
\limsup_{T\to\infty} \Pr\left(T^{1/2}\|\widehat{\theta}_T - \theta_0\| > M_1\right) < \zeta. \tag{B.22}
$$

This shows that $\|\widehat{\theta}_T - \theta_0\| = O_p(T^{-1/2})$. To show (B.22), consider that

$$
\begin{aligned}
&\Pr\left(T^{1/2}\|\widehat{\theta}_T - \theta_0\| > M_1\right) \\
&\leq \Pr\left(T^{1/2}\|\widehat{\theta}_T - \theta_0\| > M_1,\ o_p(1) \geq -eig_{\min}(\Upsilon)/2\right) + \Pr\left(o_p(1) < -eig_{\min}(\Upsilon)/2\right) \\
&\leq \Pr\left(T(eig_{\min}(\Upsilon) + o_p(1))\|\widehat{\theta}_T - \theta_0\|^2 > \frac{eig_{\min}(\Upsilon)M_1^2}{2},\ T^{1/2}\|\widehat{\theta}_T - \theta_0\| > M_1\right) + o(1) \\
&\leq \Pr\left(T(eig_{\min}(\Upsilon) + o_p(1))\|\widehat{\theta}_T - \theta_0\|^2 > \frac{eig_{\min}(\Upsilon)M_1^2}{2},\ T^{1/2}\|\widehat{\theta}_T - \theta_0\| > M_1,\ O_p(1) \geq -M_2\right) \\
&\qquad + \Pr\left(O_p(1) < -M_2\right) + o(1) \\
&\leq \Pr\left(T(eig_{\min}(\Upsilon) + o_p(1))\|\widehat{\theta}_T - \theta_0\|^2 + O_p(T^{1/2})\|\widehat{\theta}_T - \theta_0\| > \frac{eig_{\min}(\Upsilon)M_1^2}{2} - M_1M_2\right) \\
&\qquad + \Pr\left(O_p(1) < -M_2\right) + o(1) \\
&\leq \Pr\left(O_p(1) > \frac{eig_{\min}(\Upsilon)M_1^2}{2} - M_1M_2\right) + \Pr\left(O_p(1) < -M_2\right) + o(1),
\end{aligned}
$$

where the last inequality holds by (B.21), and the different $O_p(1)$ terms appearing above are not necessarily the same ones. Fix M_2 at a value such that the limsup of the second term in the last line is less than $\zeta/2$. Note that $\frac{eig_{\min}(\Upsilon)M_1^2}{2} - M_1M_2$ can be made arbitrarily large by increasing M_1 (by Assumption 10(c), $eig_{\min}(\Upsilon) > 0$). Thus, we can choose a M_1 large enough so that the limsup of the first term in the last line is also less than $\zeta/2$. Therefore, a large enough M_1 exists such that (B.22) holds. □

Proof of Lemma 1. Note that

$$\widehat{Q}_T(\theta_T) - \widehat{Q}_{0,T}(\theta_T) = \sum_{g \in \mathcal{G} \setminus \mathcal{G}_0} \mu(g)[\bar{m}_T^u(\theta_T, g)]_-^2 + \sum_{g \in \mathcal{G} \setminus \mathcal{G}_0} \mu(g)[\bar{m}_T^\ell(\theta_T, g)]_+^2.$$

Thus, it suffices to show that

$$\sum_{g \in \mathcal{G} \setminus \mathcal{G}_0} \mu(g)[\bar{m}_T^u(\theta_T, g)]_-^2 = o_p(T^{-1}), \text{ and} \tag{B.23}$$

$$\sum_{g \in \mathcal{G} \setminus \mathcal{G}_0} \mu(g)[\bar{m}_T^\ell(\theta_T, g)]_+^2 = o_p(T^{-1}). \tag{B.24}$$

We separate the two cases, one where Assumption 5 is satisfied and the other where Assumption 6 is satisfied.

Case 1: Assumption 5 is satisfied. In this case, arguments for (B.23) and (B.24) are analogous. Thus, we give the detailed proof for (B.23) only. First consider that

$$\sum_{g \in \mathcal{G} \setminus \mathcal{G}_0} \mu(g)[\bar{m}_T^u(\theta_T, g)]_-^2 \leq \sum_{g \in \mathcal{G} \setminus \mathcal{G}_0} \mu(g)[A_T(g) + B_T(g) + C_T(g)]_-^2,$$

where

$$\begin{aligned} A_T(g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t, \lambda_T) - x_{jt}'\beta_T)g(z_{jt}) \\ B_T(g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\hat{\delta}_{jt}(\tilde{s}_t, \lambda_T) - \hat{\delta}_{jt}(\pi_t, \lambda_T))g(z_{jt}) \\ C_T(g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt} + \underline{\iota}_u/n_t) - \log(\pi_{jt}))g(z_{jt}). \end{aligned} \tag{B.25}$$

The inequality holds because $\iota_u \geq \underline{\iota}_u$ (Assumption 5(b)). For $A_T(g)$, consider that

$$A_T(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt}) + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t, \lambda_T) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt}) - \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}x_{jt}(\beta_T - \beta_0)g(z_{jt}).$$

Equation (B.8) in the proof of Theorem 2 implies that

$$\sup_{g \in \mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt})\right| = O_p(T^{-1/2}). \tag{B.26}$$

Also,

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t,\lambda_T)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt})\right| \leq \sup_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}|(\delta_{jt}(\pi_t,\lambda_T)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt})|$$
$$\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}|\delta_{jt}(\pi_t,\lambda_T)-\delta_{jt}(\pi_t,\lambda_0)|$$
$$= O_p(\log(T))\|\lambda_T-\lambda_0\|, \tag{B.27}$$

where the second inequality holds because $g(z_{jt})\in(0,1)$, the first equality holds by Assumption 9(c). Moreover,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}x_{jt}(\beta_T-\beta_0)g(z_{jt}) \leq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\|x_{jt}\|\,\|\beta_T-\beta_0\| = O_p(1)\,\|\beta_T-\beta_0\|\,. \tag{B.28}$$

Therefore, combining (B.26), (B.27), (B.28) and $\|\theta_T-\theta_0\|=O_p(T^{-1/2})$, we have

$$\sup_{g\in\mathcal{G}}|A_T(g)| = O_p(\log(T)T^{-1/2}). \tag{B.29}$$

Now consider $B_T(g)$. Let $B_T^0(g)=\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\hat{\delta}_{jt}(\tilde{s}_t,\lambda_0)-\hat{\delta}_{jt}(\pi_t,\lambda_0))g(z_{jt})$. Consider that

$$\sup_{g\in\mathcal{G}}|B_T(g)-B_T^0(g)| \leq \sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t,\lambda_T)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt})\right|$$
$$+\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t,\lambda_T)-\delta_{jt}(\tilde{s}_t,\lambda_0))g(z_{jt})\right|.$$

The first summand is less than or equal to $O_p(\log(T))\|\lambda_T-\lambda_0\|$ by (B.27). The second summand is also less than or equal to $O_p(\log(T))\|\lambda_T-\lambda_0\|$ due to the same arguments as those for (B.27). Those combined with $\|\theta_T-\theta_0\|=O_p(T^{-1/2})$ shows that:

$$\sup_{g\in\mathcal{G}}|B_T(g)-B_T^0(g)| = O_p(\log(T)T^{-1/2}). \tag{B.30}$$

For $B_T^0(g)$, consider that

$$B_T^0(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\hat{\delta}_{jt}(\tilde{s}_t,\lambda_0)-\hat{\delta}_{jt}(\pi_t,\lambda_0))g(z_{jt})$$
$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}g(z_{jt})\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}(\tilde{s}_t-s_t)$$

$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} g(z_{jt})\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}(s_t-\pi_t)$$
$$+\frac{1}{2T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} g(z_{jt})(\tilde{s}_t-\pi_t)'\frac{\partial^2\hat{\delta}_{jt}(\tilde{\pi}_t,\lambda_0))}{\partial\pi\partial\pi'}(\tilde{s}_t-\pi_t), \tag{B.31}$$

where $\tilde{\pi}_t$ is a point on the line segment connecting $\tilde{s}_t$ and π_t. For the first summand, note that, by the Cauchy-Schwartz inequality and $g(z)\in[0,1]$, its absolute value is less than or equal to

$$\left(\sup_{t=1,\ldots,T} n_t\|\tilde{s}_t-s_t\|\right)\left(\frac{1}{T\underline{n}_T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\left\|\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}\right\|\right)=O_p(1)\underline{n}_T^{-1}O_p(1)=o_p(T^{-1/2}),$$

where the first equality holds by Assumption 4,

$$E\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\left\|\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}\right\|\right)\leq\bar{J}\sup_{j,t}E\left\|\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}\right\|<\infty$$

(by Assumption 10(b)), and the Chebyschev inequality, and the second equality holds by Assumption 9(d). For the second summand of (B.31), we can apply Lemma 4 and get

$$\begin{aligned}
&E\left[\sup_{g\in\mathcal{G}}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} g(z_{jt})\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}(s_t-\pi_t)\right)^2\right]\\
&\leq\frac{C\bar{J}}{T^2}\sum_{t=1}^{T}\sum_{j=1}^{J_t}E\left(\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}(s_t-\pi_t)\right)^2\\
&=\frac{C\bar{J}}{T^2}\sum_{t=1}^{T}\sum_{j=1}^{J_t}E\left(\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}\frac{diag(\pi_t)-\pi_t\pi_t'}{n_t}\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}\right)\\
&\leq\frac{C\bar{J}}{\underline{n}_T T^2}\sum_{t=1}^{T}\sum_{j=1}^{J_t}E\left(\left\|\frac{\partial\hat{\delta}_{jt}(\pi_t,\lambda_0))}{\partial\pi'}\right\|^2\right)\\
&=O(\underline{n}_T^{-1}T^{-1})\\
&=o(T^{-1}),
\end{aligned}$$

where the first equality holds by $E[(s_t-\pi_t)(s_t-\pi_t)']=\frac{diag(\pi_t)-\pi_t\pi_t'}{n_t}$ which holds under Assumption 7(b), the second inequality holds because $diag(\pi_t)-\pi_t\pi_t'$ is positive semi-definite and its largest eigenvalue of does not exceed the highest π_{jt} which does not exceed 1 and because $n_t\geq\underline{n}_T$ for all $t=1,\ldots,T$, the second equality holds by Assumption 10(b), and the last equality holds by Assumption 9(d). Therefore, the Markov inequality applies and shows that the second summand

of (B.31) is $o_p(T^{-1/2})$ uniformly over $g \in \mathcal{G}$. For the third summand of (B.31), consider that

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} \left| \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t} g(z_{jt})(\tilde{s}_t - \pi_t)' \frac{\partial^2 \hat{\delta}_{jt}(\tilde{\pi}_t, \lambda_0))}{\partial\pi\partial\pi'}(\tilde{s}_t - \pi_t) \right| \\
&\leq_{w.p.a.1.} \sup_{j,t} \sup_{\pi:\|\pi-\pi_t\|\leq c} \left\| \frac{\partial^2 \hat{\delta}_{jt}(\tilde{\pi}_t, \lambda_0))}{\partial\pi\partial\pi'} \right\| \bar{J} T^{-1} \sum_{t=1}^{T} (\tilde{s}_t - \pi_t)'(\tilde{s}_t - \pi_t) \\
&= O_p(1) 2 \left[T^{-1}\sum_{t=1}^{T} \|\tilde{s}_t - s_t\|^2 + T^{-1}\sum_{t=1}^{T} \|s_t - \pi_t\|^2 \right] \\
&= O_p(1) O_p(\underline{n}_T^{-1}) + O_p(1) T^{-1} \sum_{t=1}^{T} \|s_t - \pi_t\|^2 \\
&= O_p(1) O_p(\underline{n}_T^{-1}) + O_p(1) O_p(\underline{n}_T^{-1}) \\
&= o_p(T^{-1/2}),
\end{aligned}$$

where the first inequality holds because $\sup_t \|\tilde{\pi}_t - \pi_t\| \leq c$ w.p.a.1. by Assumptions 4 and 7(g) and also because $g(z) \in [0,1]$, the first equality holds by Assumption 10(b), the second equality holds by Assumption 4, the third equality holds by Chebyshev inequality and $E\|s_t - \pi_t\|^2 = E\sum_{j=1}^{J_t} \pi_{jt}(1-\pi_{jt})/n_t \leq \underline{n}_T^{-1}$, and the last equality holds by Assumption 9(d). Combining the arguments for all the three summands in (B.31), we have

$$\sup_{g\in\mathcal{G}} \left| B_T^0(g) \right| = o_p(T^{-1/2}). \tag{B.32}$$

This and (B.30) together imply that

$$\sup_{g\in\mathcal{G}} |B_T(g)| = o_p(T^{-1/2}). \tag{B.33}$$

Next consider $C_T(g)$. Using the moment bound derived in (B.9) in the proof of Theorem 2 and the Markov inequality, we can derive

$$\sup_{g\in\mathcal{G}} |C_T(g) - E[C_T(g)]| = O_p\left(\frac{\log \overline{n}_T}{T^{1/2}}\right) = O_p\left(\frac{\log T}{T^{1/2}}\right), \tag{B.34}$$

where the second equality holds by $\overline{n}_T T^{-2} \to_p 0$ (Assumption 9(d)).

Let $r_T(g)$ denote $A_T(g) + B_T(g) + C_T(g) - E[C_T(g)]$. Then $\bar{m}_T^u(\theta_T, g) \geq r_T(g) + E[C_T(g)]$. And by equations (B.29), (B.33), and (B.34), we have

$$\sup_{g\in\mathcal{G}} |r_T(g)| = O_p(T^{-1/2} \log T). \tag{B.35}$$

For a sequence c_T such that $T^{-1/2}\log T = o(c_T)$, consider:

$$\begin{aligned}\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]>c_T}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2 &\le \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]>c_T}\mu(g)[r_T(g)+c_T]_-^2\\ &\le \sup_{g\in\mathcal{G}}[r_T(g)+c_T]_-^2\\ &= [o_p(c_T)+c_T]_-^2\\ &=_{w.p.a.1} 0, \end{aligned} \tag{B.36}$$

where the first inequality holds because $[\cdot]_-^2$ is nonincreasing, the second inequality holds because $\mu(g)$ is a probability mass function, the first equality holds by (B.35). Thus, the expression $\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]>c_T}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2$ converges in probability to zero at arbitrary rate. Further restrict c_T so that $c_T = o((\log T)^{-2/\eta})$. This is possible because for any finite $\eta > 0$, $\log(T)^{1+2/\eta} = o(T^{1/2})$. Also consider

$$\begin{aligned}\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]\le c_T}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2 &\le \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]\le c_T}\mu(g)[r_T(g)]_-^2\\ &\le \sup_{g\in\mathcal{G}}|r_T(g)|^2 \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]\le c_T}\mu(g)\\ &= O_p\left(T^{-1}(\log T)^2\right)c_T^\eta\\ &= o_p(T^{-1}), \end{aligned} \tag{B.37}$$

where the first inequality holds because $\bar{m}_T^u(\theta_T,g) = r_T(g) + E[C_T(g)]$ and

$$E[C_T(g)] = T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}E[\log(s_{jt}+\iota_u/n_t) - \log(\pi_{jt})] \ge 0$$

by the definition of ι_u, and the first equality holds by the first part of Assumption 10(a). Therefore, we have

$$\begin{aligned}\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2 &= \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]>c_n}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2 + \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0:E[C_T(g)]\le c_n}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2\\ &= o_p(T^{-1}). \end{aligned} \tag{B.38}$$

Case 2: Assumption 6 is satisfied. We prove (B.23) first. Observe that

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)[\bar{m}_T^u(\theta_T,g)]_-^2 = \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)[A_T(g)+\Delta_T(g)+S_T(g)]_-^2,$$

where

$$A_T(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t, \lambda_T) - x'_{jt}\beta_T)g(z_{jt})$$
$$\Delta_T(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t, \lambda_T) - \delta_{jt}(\pi_t, \lambda_T))g(z_{jt})$$
$$S_T(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt} + \iota_u/n_t) - \log(\tilde{s}_{jt}))g(z_{jt}). \tag{B.39}$$

The same arguments showing (B.29) in Case 1 still applies in Case 2 since neither Assumption 5 or Assumption 6 is involved. Thus, (B.29) holds. For $\Delta_T(g)$, the same arguments as those for (B.30) shows that

$$\sup_{g\in\mathcal{G}\setminus\mathcal{G}_0}\left|\Delta_T(g) - \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt})\right| = O_p(\log(T))\|\lambda_T - \lambda_0\|. \tag{B.40}$$

Equation (B.13) in Case 2 of the proof of Theorem 2 shows that

$$\begin{aligned}&E\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt}) - E[(\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt})]\right|^2\\ =&O\left(\frac{\log(\overline{n}_T)^2}{T}\right).\end{aligned}$$

Thus, by the Markov inequality,

$$\begin{aligned}&\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt}) - E[(\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt})]\right|\\ =&O_p\left(\frac{\log(\overline{n}_T)}{T^{1/2}}\right)\\ =&O_p(\log(T)T^{-1/2}).\end{aligned} \tag{B.41}$$

where the second equality holds by $\overline{n}_T T^{-2} \to_p 0$ (Assumption 9(d)). By Assumption 6(b), $E[(\delta_{jt}(\tilde{s}_t, \lambda_0) - \delta_{jt}(\pi_t, \lambda_0))g(z_{jt}) \geq 0$. This combined with (B.40), (B.41), and $\|\widehat{\theta}_T - \theta_0\| = O_p(T^{-1/2})$ implies that

$$\inf_{g\in\mathcal{G}}\Delta_T(g) \geq O_p((\log(T)T^{-1/2}). \tag{B.42}$$

For $S_T(g)$, note that

$$S_T(g) \geq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(s_{jt}+\iota_u/n_t)^{-1}((s_{jt}+\iota_u/n_t)-(\tilde{s}_{jt}))g(z_{jt})$$
$$= (\iota_u - 1)\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\frac{g(z_{jt})}{n_t s_{jt}+\iota_u}.$$

Applying Lemma 4 and using the fact that $E[(n_t s_{jt}+\iota_u)^{-2}] \leq \iota_u^{-2}$, we have

$$E\sup_{g\in\mathcal{G}}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\frac{g(z_{jt})}{n_t s_{jt}+\iota_u} - E\left[\frac{g(z_{jt})}{n_t s_{jt}+\iota_u}\right]\right)^2 = O(T^{-1}).$$

Then by the Markov inequality we have

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\frac{g(z_{jt})}{n_t s_{jt}+\iota_u} - E\left[\frac{g(z_{jt})}{n_t s_{jt}+\iota_u}\right]\right| = O_p(T^{-1/2}).$$

Thus we have

$$S_T(g) \geq O_p(T^{-1/2}) + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}E\left[\frac{g(z_{jt})}{n_t s_{jt}+\iota_u}\right]. \tag{B.43}$$

Using (B.29), (B.42), (B.43), and the third part of Assumption (10)(a), we can apply the same arguments as those for (B.38) (from (B.36) to (B.38)) to conclude that (B.23) holds.

Finally we prove (B.24) for Case 2. Note that

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)[\bar{m}_T^\ell(\theta_T,g)]_+^2 \leq \sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)[A_T(g)+B_T(g)+C_T^\ell(g)]_+^2,$$

where

$$A_T(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t,\lambda_T)-x_{jt}'\beta_T)g(z_{jt})$$
$$B_T(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\hat{\delta}_{jt}(\tilde{s}_t,\lambda_T)-\hat{\delta}_{jt}(\pi_t,\lambda_T))g(z_{jt})$$
$$C_T^\ell(g) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt}+\bar{\iota}_\ell/n_t)-\log(\pi_{jt}))g(z_{jt}). \tag{B.44}$$

The same arguments showing (B.29) in Case 1 still applies in Case 2 since neither Assumption 5 or Assumption 6 is involved. Thus, (B.29) holds. For $B_T(g)$, the same arguments for (B.30) in Case 1 still applies here as well. Thus, (B.30) holds, and we only need to study $B_T^0(g)$ to understand the

behavior of $B_T(g)$. Note that

$$\begin{aligned}B_T^0(g) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\tilde{s}_t,\lambda_0)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt}) - E[(\delta_{jt}(\tilde{s}_t,\lambda_0)-\delta_{jt}(\pi_t,\lambda_0))g(z_{jt})] \\ &- \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(\tilde{s}_t)-\log(\pi_t))g(z_{jt}) - E[(\log(\tilde{s}_t)-\log(\pi_t))g(z_{jt})] \\ &+ \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J_t}E[(\hat{\delta}_{jt}(\tilde{s}_t,\lambda_0)-\hat{\delta}_{jt}(\pi_t,\lambda_0))g(z_{jt})].\end{aligned}$$

Equation (B.41) shows that the first summand is $O_p(\log(T)T^{-1/2})$ uniformly over $g \in \mathcal{G}$, Equation (B.9) and Markov inequality combined show that the second summand is $O_p(\log(T)T^{-1/2})$ uniformly over $g \in \mathcal{G}$. The third summand is non-positive by Assumption 6(a). Therefore

$$\sup_{g\in\mathcal{G}} B_T(g) \leq O_p(\log(T)T^{-1/2}). \tag{B.45}$$

The same arguments as those for the second summand above shows that $\sup_{g\in\mathcal{G}}|C_T^\ell(g)-E[C_T^\ell(g)]| = O_p(\log(T)T^{-1/2})$. Using this, (B.29), (B.45), and the second part of Assumption (10)(a), we can apply the same arguments as those for ((B.38)) (from ((B.36)) to ((B.38))) to conclude that (B.24) holds. □

Proof of Lemma 2. (a) By equation (B.4) in the proof of Theorem 1, we have

$$\begin{aligned}&\sup_{\theta\in B_c(\theta_0)}\left|\sqrt{\widehat{Q}_{0,T}(\theta)}-\sqrt{\widehat{Q}_T^*(\theta)}\right| \\ &\leq \sqrt{\sup_{\theta\in B_c(\theta_0)}\sup_{g\in\mathcal{G}_0}|\bar{m}_T^u(\theta,g)-\bar{m}_T(\theta,g)|^2 + \sup_{\theta\in B_c(\theta_0)}\sup_{g\in\mathcal{G}_0}|\bar{m}_T^\ell(\theta,g)-\bar{m}_T(\theta,g)|^2}.\end{aligned} \tag{B.46}$$

Now note that

$$\begin{aligned}\bar{m}_T^u(\theta,g)-\bar{m}_T(\theta,g) &= T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}^u(s_t,\lambda)-\delta_{jt}(\pi_t,\lambda))g(z_{jt}) \\ &= T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt}+\iota_u/n_t)-\log(s_{jt}))g(z_{jt}) \\ &+ T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}\delta_{jt}(\tilde{s}_t,\lambda)-\delta_{jt}(\pi_t,\lambda))g(z_{jt}).\end{aligned} \tag{B.47}$$

For the first summand, consider that

$$\sup_{g\in\mathcal{G}_0}\left|T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}(\log(s_{jt}+\iota_u/n_t)-\log(s_{jt}))g(z_{jt})\right| = T^{-1}\iota_u\sum_{t=1}^{T}\sum_{j=1}^{J_t}((s_{jt}+\tilde{\iota}/n_t)^{-1}n_t^{-1}$$
$$\leq \underline{n}_T^{-1}\bar{J}\iota_u \sup_{j,t:z_{jt}\in\mathcal{Z}_0} s_{jt}^{-1}$$
$$= O_p(\underline{n}_T^{-1}) = o_p(T^{-1/2}), \tag{B.48}$$

where the first equality holds with $\tilde{\iota}\in[0,\iota_u]$ by mean-value expansion, the inequality holds by the definition of $\mathcal{G}_0$, the second equality holds because s_{jt} is bounded away from zero by Assumptions 1(a) and 7(g). For the second summand in (B.47), we can apply the same arguments as those for (B.32) to show that this second summand is $o_p(T^{-1/2})$ with the following adjustment: (1) Replace $\mathcal{G}$ by $\mathcal{G}_0$, and (2) realize that the second derivative part of Assumption 10(b) in the case of either Assumpiton 5 or Assumption 6 is sufficient for the current purpose. Therefore, we have

$$\sup_{\theta\in B_c(\theta_0)}\sup_{g\in\mathcal{G}_0}|\bar{m}_T^u(\theta,g)-\bar{m}_T(\theta,g)| = o_p(T^{-1/2}).$$

Analogous arguments can be used to show that $\sup_{\theta\in B_c(\theta_0)}\sup_{g\in\mathcal{G}_0}|\bar{m}_T^\ell(\theta,g)-\bar{m}_T(\theta,g)| = o_p(T^{-1/2})$. That concludes the proof. □

(b) Recall that $\widehat{Q}_T^*(\theta_0)=\sum_{g\in\mathcal{G}_0}\mu(g)(\bar{m}_T(\theta_0,g))^2$, and note that

$$\bar{m}_T(\theta_0,g)=\frac{1}{T}\sum_{T=1}^{T}\sum_{j=1}^{J_t}(\delta_{jt}(\pi_t,\lambda_0)-x_{jt}'\beta_0)g(z_{jt})=\frac{1}{T}\sum_{T=1}^{T}\sum_{j=1}^{J_t}\xi_{jt}g(z_{jt}).$$

Then by equation(B.8) in the proof of Theorem 2, we have

$$\sup_{g\in\mathcal{G}_0}|\bar{m}_T(\theta_0,g)| = O_p(T^{-1/2}). \tag{B.49}$$

This is sufficient for part (b) to hold.

Proof of Lemma 3. (a) First consider that

$$\begin{aligned}
&\bar{m}_T(\theta_T,g)-\bar{m}_T(\theta_0,g)\\
&=T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}g(z_{jt})[\delta_{jt}(\pi_t,\lambda_T)-\delta_{jt}(\pi_t,\lambda_0)+x_{jt}(\beta_T-\beta_0)]\\
&=T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}g(z_{jt})\partial m_{jt}(\lambda_0)'(\theta_T-\theta_0)+T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}g(z_{jt})(\lambda_T-\lambda_0)'\frac{\partial^2\delta_{jt}(\pi_t,\tilde{\lambda})}{\partial\lambda\partial\lambda'}(\lambda_T-\lambda_0)/2\\
&=\widehat{\Gamma}_T(g)'(\theta_T-\theta_0)+(\lambda_T-\lambda_0)'D_T(g)(\lambda_T-\lambda_0),
\end{aligned}$$

where $\tilde{\lambda}$ is a point on the line segment connecting λ_T and λ_0, and

$$D_T(g) = (2T)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} g(z_{jt}) \frac{\partial^2 \delta_{jt}(\pi_t, \tilde{\lambda})}{\partial \lambda \partial \lambda'}.$$

Thus, we have

$$\begin{aligned}
&\widehat{Q}_T^*(\theta_T) - \widehat{Q}_T^*(\theta_0) \\
&= \sum_{g \in \mathcal{G}_0} \mu(g)(\bar{m}_T(\theta_T, g) - \bar{m}_T(\theta_0, g))^2 + 2 \sum_{g \in \mathcal{G}_0} \mu(g) \bar{m}_T(\theta_0, g)(\bar{m}_T(\theta_T, g) - \bar{m}_T(\theta_0, g)) \qquad \text{(B.50)} \\
&= (\theta_T - \theta_0) \sum_{g \in \mathcal{G}_0} \mu(g) \widehat{\Gamma}_T(g) \widehat{\Gamma}_T(g)' (\theta_T - \theta_0) \\
&\quad + 2 \sum_{g \in \mathcal{G}_0} \mu(g)(\lambda_T - \lambda_0)' D_T(g)(\lambda_T - \lambda_0) \widehat{\Gamma}_T(g)' (\theta_T - \theta_0) \\
&\quad + \sum_{g \in \mathcal{G}_0} \mu(g)\{(\lambda_T - \lambda_0)' D_T(g)(\lambda_T - \lambda_0)\}^2 \\
&\quad + 2 \sum_{g \in \mathcal{G}_0} \mu(g) \bar{m}_T(\theta_0, g) \widehat{\Gamma}_T(g)' (\theta_T - \theta_0) \\
&\quad + 2 \sum_{g \in \mathcal{G}_0} \mu(g) \bar{m}_T(\theta_0, g)(\lambda_T - \lambda_0)' D_T(g)(\lambda_T - \lambda_0). \qquad \text{(B.51)}
\end{aligned}$$

Since $\tilde{\lambda} \in B_c(\lambda_0)$ whenever $\lambda_T \in B_c(\lambda_0)$ (which holds with probability approaching one because $\|\lambda_T - \lambda_0\| \to_p 0$), we have for any $g \in \mathcal{G}_0$,

$$\sup_{g \in \mathcal{G}_0} \|D_T(g)\| \leq_{w.p.a.1} \bar{J} \sup_{j,t} \sup_{\lambda : \|\lambda - \lambda_0\| \leq c} \left\| \frac{\partial^2 \delta_{jt}(\pi_t, \tilde{\lambda})}{\partial \lambda \partial \lambda'} 1(z_{jt} \in \mathcal{Z}_0) \right\| = O_p(1), \qquad \text{(B.52)}$$

where the first inequality holds because $0 \leq g(z) \leq 1$ and $\tilde{\lambda}$ is in a c-neighborhood of λ_0 with probability approaching one (by $\|\theta_T - \theta_0\| = o_p(1)$), and the equality holds by Assumption 10(c). This combined with $\|\theta_T - \theta_0\| = o_p(1)$ implies that

$$\sum_{g \in \mathcal{G}_0} \mu(g)\{(\lambda_T - \lambda_0)' D_T(g)(\lambda_T - \lambda_0)\}^2 \leq \sup_{g \in \mathcal{G}_0} \|D_T(g)\|^2 \|\theta_T - \theta_0\|^4 = o_p(1)\|\theta_T - \theta_0\|^2.$$

Also, by the first part of Assumption 10(c), we have $\sup_{g \in \mathcal{G}_0} \|\widehat{\Gamma}_T(g)\| = O_p(1)$. This combined with (B.52) and $\|\theta_T - \theta_0\| = o_p(1)$ implies that

$$\begin{aligned}
\left| \sum_{g \in \mathcal{G}_0} \mu(g)(\lambda_T - \lambda_0)' D_T(g)(\lambda_T - \lambda_0) \widehat{\Gamma}_T(g)' (\theta_T - \theta_0) \right| &\leq \|\theta_T - \theta_0\|^3 \sup_{g \in \mathcal{G}_0} \|D_T(g)\| \|\widehat{\Gamma}_T(g)\| \\
&= o_p(1)\|\theta_T - \theta_0\|^2.
\end{aligned}$$

Next apply Lemma 4 with $w_{jt} = \xi_{jt}$ and we get

$$\begin{aligned} E \sup_{g \in \mathcal{G}_0} (\bar{m}_T(\theta_0, g))^2 &= E \sup_{g \in \mathcal{G}_0} \left(T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \xi_{jt} g(z_{jt}) \right)^2 \\ &\leq C \bar{J} T^{-2} \sum_{t=1}^{T} \sum_{j=1}^{J_t} E[\xi_{jt}^2 1(z_{jt} \in \mathcal{Z}_0)] \\ &= O(T^{-1}), \end{aligned}$$

where the second equality holds by Assumptions 7(d) and (f). Therefore,

$$\sup_{g \in \mathcal{G}_0} |\bar{m}_T(\theta_0, g)| = O_p(T^{-1/2}). \tag{B.53}$$

This combined with (B.52) implies that

$$\sum_{g \in \mathcal{G}_0} \mu(g) \bar{m}_T(\theta_0, g)(\lambda_T - \lambda_0)' D_T(g)(\lambda_T - \lambda_0) = O_p(T^{-1/2}) \|\theta_T - \theta_0\|^2.$$

Therefore, part (a) holds.

(b) Apply Lemma 4 with w_{jt} being an element of the random vector $\partial m_{jt}(\lambda_0)$, do so for every element of $\partial m_{jt}(\lambda_0)$, and we get

$$\begin{aligned} E \sup_{g \in \mathcal{G}_0} \left\| \widehat{\Gamma}_T(g) - \Gamma_T(g) \right\|^2 &\leq C \bar{J} T^{-2} \sum_{t=1}^{T} \sum_{j=1}^{J_t} E[\|\partial m_{jt}(\lambda_0)\|^2 1(z_{jt} \in \mathcal{Z}_0)]. \\ &= O(T^{-1}). \end{aligned}$$

The equality is implied by Assumptions 7(d) and 10(c). Thus, we have

$$\sup_{g \in \mathcal{G}_0} \left\| \widehat{\Gamma}_T(g) - \Gamma_T(g) \right\| = O_p(T^{-1/2}). \tag{B.54}$$

Assumption 9(c) implies that

$$\begin{aligned} \sup_{g \in \mathcal{G}_0} \|\Gamma_T(g)\| &\leq \sup_{g \in \mathcal{G}_0} T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} E[\|\partial m_{jt}(\lambda_0) g(z_{jt})\|] \\ &\leq \sup_{g \in \mathcal{G}_0} T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} E[\|\partial m_{jt}(\lambda_0)\| 1(z_{jt} \in \mathcal{Z}_0)] \\ &= O(1). \end{aligned} \tag{B.55}$$

This and (B.54) together imply that

$$\widehat{\Upsilon}_T = \sum_{g\in\mathcal{G}_0} \mu(g)\widehat{\Gamma}_T(g)\widehat{\Gamma}_T(g)' = o_p(1) + \sum_{g\in\mathcal{G}_0} \mu(g)\Gamma_T(g)\Gamma_T(g)' \to_p \Upsilon,$$

where the convergence holds by Assumption 10(d).

For W_n, first consider the derivation

$$\begin{aligned}\left| T^{1/2} \sum_{g\in\mathcal{G}_0} \mu(g)\bar{m}_T(\theta_0, g)(\widehat{\Gamma}_T(g) - \Gamma_T(g)) \right| &\leq \sup_{g\in\mathcal{G}_0} |\bar{m}_T(\theta_0, g)| \sup_{g\in\mathcal{G}_0} T^{1/2}\|\widehat{\Gamma}_T(g) - \Gamma_T(g)\| \\ &= O_p(T^{-1/2}) = o_p(1),\end{aligned}$$

by equations (B.53) and (B.54). Thus,

$$\begin{aligned}T^{1/2}W_n &= o_p(1) + T^{1/2} \sum_{g\in\mathcal{G}_0} \mu(g)\bar{m}_T(\theta_0, g)\Gamma_T(g) \\ &= o_p(1) + T^{-1/2} \sum_{t=1}^{T} v_t,\end{aligned}$$

where $v_t = \sum_{j=1}^{J_t} \left[\xi_{jt} \left(\sum_{g\in\mathcal{G}_0} \mu(g)g(z_{jt})\Gamma_T(g)\right)\right]$. Observe that $\{v_t\}_{t=1}^T$ is independent across t by Assumption 7(e),

$$E[v_t] = E\sum_{j=1}^{J_t} \left[E[\xi_{jt}|z_{jt}] \left(\sum_{g\in\mathcal{G}_0} \mu(g)g(z_{jt})\Gamma_T(g) \right) \right] = 0$$

$$T^{-1}\sum_{t=1}^{T} E[v_t v_t'] = T^{-1}\sum_{t=1}^{T} \sum_{g,g^*\in\mathcal{G}_0} Cov\left(\sum_{j=1}^{J_t} \xi_{jt}g(z_{jt}), \sum_{j=1}^{J_t} \xi_{jt}g^*(z_{jt}) \right) \Gamma_T(g)\Gamma_T(g)'\mu(g)\mu(g^*) \to V,$$

by Assumptions 7(c) and 10(e), and for the c in Assumption 7(f),

$$\begin{aligned}E(\|v_t\|^{2+c}) &\leq \bar{J}^{1+c} E|\xi_{jt}|^{2+c} \sup_{g\in\mathcal{G}_0} \|\Gamma_T(g)\|^{2+c} \\ &= O(1),\end{aligned}$$

by Assumptions 7(d) and (f) and equation (B.55) above. Therefore, we can apply the Lindeberg central limit theorem and conclude $T^{-1/2}\sum_{t=1}^{T} v_t \to_d N(0, V)$. Therefore,

$$T^{1/2}W_n \to_d N(0, V).$$

□

B.4 Auxiliary Lemmas

The following lemma establishes a maximal inequality for certain empirical processes indexed by g in a subset of $\mathcal{G}$.

Lemma 4. *Let $\{z_{jt} : j = 1, \ldots, J_t, t = 1, \ldots, T\}_{T \geq 1}$ be an array of random vectors, where $\max_{t=1}^{T} J_t \leq \bar{J} < \infty$. Let $\mathcal{G}$ be the set of instrumental functions defined in (4.9). Let $\mathcal{Z}^*$ be a subset of $supp(z_{jt})$ and let $\mathcal{G}^*$ be a subset of $\mathcal{G}$ for which $g(z) = 0$ for all $z \notin \mathcal{Z}^*$for all $g \in \mathcal{G}^*$. Let $\{w_{jt} : j = 1, \ldots, J_t, t = 1, \ldots, T\}_{T \geq 1}$ be an array of random variables such that $\sum_{t=1}^{T} \sum_{j=1}^{J_t} E[w_{jt}^2 1(z_{jt} \in \mathcal{Z}^*)] \leq a(T)$ for a function $a(T)$ of T for all T. Let $w_t = (w_{1t}, \ldots, w_{J_t t})'$ and $z_t = (z_{1t}, \ldots, z_{J_t t})'$. Suppose that (w_t, z_t) is independent across t. Then*

$$E \sup_{g \in \mathcal{G}^*} \left(T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (w_{jt} g(z_{jt}) - E[w_{jt} g(z_{jt})]) \right)^2 \leq C \bar{J} a(T)/T^2,$$

for some constant $C > 0$.

Proof. First observe that $\sum_{j=1}^{J_t} w_{jt} g(z_{jt})$ can be written as $f_t(g) := \sum_{j=1}^{\bar{J}} w_{jt} 1(j \leq J_t) g(z_{jt})$. Observe that the triangular array of random processes $\{g(z_{jt}) : g \in \mathcal{G}^* : t = 1, \ldots, T\}_{T \geq 1}$ is manageable with respect to the envelope $\mathbf{1}_T$ for all j in the sense of Pollard (1990) because $\mathcal{G}$ is the collection of indicator functions for a Vapnik-Cervonenkis class of sets. Then by parts (a) and (c) of Lemma E1 in Andrews and Shi (2013), we have that the triangular array $\{f_t(g) : g \in \mathcal{G}^* : t = 1, \ldots, T; T \geq 1\}$ is manageable with respect to the envelope function $F_T = (F_{T1}, \ldots, F_{TT})$ where $F_{Tt} = \sum_{j=1}^{\bar{J}} 1(j \leq J_t, z_{jt} \in \mathcal{Z}^*)|w_{jt}| \equiv \sum_{j=1}^{J_t} |w_{jt}| 1(z_{jt} \in \mathcal{Z}^*)$. Therefore, by the maximal inequality (7.10) in Pollard (1990), we have, for some constant $C > 0$,

$$\begin{aligned} E \sup_{g \in \mathcal{G}^*} \left| \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (w_{jt} g(z_{jt}) - E[w_{jt} g(z_{jt})]) \right|^2 &\leq \frac{C}{T^2} \sum_{t=1}^{T} E[(F_{Tt})^2] \\ &\leq \frac{C \bar{J}}{T^2} \sum_{t=1}^{T} \sum_{j=1}^{J_t} E[w_{jt}^2 1(z_{jt} \in \mathcal{Z}^*)] \\ &\leq \frac{C \bar{J} a(T)}{T^2}. \end{aligned} \tag{B.56}$$

□

C Random Coefficient Logit

In this section, we prove a lemma that establishes Assumption 5 for the random coefficient logit model.

Lemma 5. *Consider the random coefficient logit model in Example 2. Assume that (i) w_{jt} is bounded, i.e. $||w_{jt}|| \leq \overline{w}$; (ii) $\sup_{\lambda \in \Lambda} \sup_{\|w\| \leq \overline{w}} \int \exp(2w'v) dF(v; \lambda) < \infty$, (iii) $\inf_{t=1,\ldots,T} \pi_{0t} \geq$*

$\underline{\varepsilon}_0 > 0$ *for all* T*, and (iv) there exists* $e_1 > 0$ *and* $0 < e_2 < \underline{\varepsilon}_0/(2\bar{J})$ *such that, the maximum eigenvalue of* $\int \tilde{\pi}_t(v)\tilde{\pi}_t(v)'dF(v;\lambda) \begin{pmatrix} \tilde{\pi}_{1t}^{-1} & 0 & \dots & 0 \\ 0 & \tilde{\pi}_{2t}^{-1} & \dots & 0 \\ \dots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \tilde{\pi}_{J_t t}^{-1} \end{pmatrix}$ *is less than* $1 - e_1$ *for all* $\lambda \in \Lambda$*, and all* $\tilde{\pi}_t$ *such that* $\|\tilde{\pi}_t - \pi_t\| < e_2$ *for all* $t = 1, \dots, T$*;* $T = 1, 2, 3, \dots$*, where*

$$\tilde{\pi}_{jt}(v) = \frac{\exp(w_{jt}'v + \delta_{jt}(\tilde{\pi}_t;\lambda))}{1 + \sum_{k=1}^{J_t} \exp(w_{kt}'v + \delta_{kt}(\tilde{\pi}_t;\lambda))}.$$

Then Assumption 5(a) *is satisfied.*

Proof. Without loss of generality, consider the derivative with respect to π_{1t}. For $j = 1, \dots, J_t$, take partial derivative with respect to π_{1t} on both sides of (5.1), and we get:

$$\begin{aligned}
&\frac{\partial \hat{\delta}_{jt}(\tilde{\pi}_t;\lambda)}{\partial \pi_{1t}} \\
&= \int \frac{\exp(w_{jt}'v)\exp(\hat{\delta}_{jt}(\tilde{\pi}_t;\lambda))}{\left(1 + \sum_{k=1}^{J_t} \exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda) + w_{kt}'v)\tilde{\pi}_{kt}\right)^2} \\
&\cdot \left(\exp(\hat{\delta}_{1t}(\tilde{\pi}_t;\lambda) + w_{1t}'v) + \sum_{k=1}^{J_t} \tilde{\pi}_{kt} \exp(w_{kt}'v)\exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)) \frac{\partial \hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)}{\partial \pi_{1t}} \right) dF(v;\lambda) \\
&= \tilde{\pi}_{1t}^{-1}\tilde{\pi}_{jt}^{-1} \int \tilde{\pi}_{jt}(v)\tilde{\pi}_{1t}(v)dF(v;\lambda) + \sum_{k=1}^{J_t} \left\{ \left[\tilde{\pi}_{jt}^{-1} \int \tilde{\pi}_{jt}(v)\tilde{\pi}_{kt}(v)dF(v;\lambda) \right] \frac{\partial \hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)}{\partial \pi_{1t}} \right\}.
\end{aligned}$$

Stacking the J_t equations in matrix form, we find that

$$H_t(\tilde{\pi}_t,\lambda)\frac{\partial \hat{\delta}_t(\tilde{\pi}_t;\lambda)}{\partial \pi_{1t}} = b_t(\tilde{\pi}_t;\lambda),$$

where

$$H_t(\tilde{\pi}_t,\lambda) = I - \int \tilde{\pi}_t(v)\tilde{\pi}_t(v)'dF(v;\lambda) \begin{pmatrix} \tilde{\pi}_{1t}^{-1} & 0 & \dots & 0 \\ 0 & \tilde{\pi}_{2t}^{-1} & \dots & 0 \\ \dots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \tilde{\pi}_{J_t t}^{-1} \end{pmatrix},$$

and

$$b_t(\tilde{\pi}_t;\lambda) = \begin{pmatrix} \tilde{\pi}_{1t}^{-2} \int \tilde{\pi}_{1t}(v)^2 dF(v;\lambda) \\ \tilde{\pi}_{1t}^{-1}\tilde{\pi}_{2t}^{-1} \int \tilde{\pi}_{1t}(v)\tilde{\pi}_{2t}(v)dF(v;\lambda) \\ \vdots \\ \tilde{\pi}_{1t}^{-1}\tilde{\pi}_{J_t t}^{-1} \int \tilde{\pi}_{1t}(v)\tilde{\pi}_{J_t t}(v)dF(v;\lambda) \end{pmatrix}.$$

By condition (iv), we have that the eigenvalues of $H_t(\tilde{\pi}_t,\lambda)$ are positive and bounded away from zero for all t, all λ and all $\tilde{\pi}_t$ in the e_2-neighborhood of π_t. Next we show that $b_t(\tilde{\pi}_t;\lambda)$ is uniformly

bounded, which will then imply that

$$\sup_{t=1,...,T;T=1,2,...} \sup_{\lambda\in\Lambda} \| \frac{\partial \hat{\delta}_t(\tilde{\pi}_t;\lambda)}{\partial \pi_{1t}} || < \infty.$$

This shows that for any consistent estimator $\hat{\pi}_t$ of π_t such that $\sup_t ||\hat{\pi}_t - \pi_t|| \to_p 0$, we have

$$\sup_t \sup_{\lambda\in\Lambda} ||\hat{\delta}_t(\hat{\pi}_t;\lambda) - \hat{\delta}_t(\pi_t;\lambda)|| \leq_{w.p.a.1} \sup_t \sup_{\lambda\in\Lambda} \sup_{||\tilde{\pi}_t - \pi_t||<e} ||\frac{\partial \hat{\delta}_t(\tilde{\pi}_t;\lambda)}{\partial \pi_{1t}}|| ||\hat{\pi}_t - \pi_t||.$$

Thus Assumption 5(a) holds.

To show that $b_t(\tilde{\pi}_t;\lambda)$ is uniformly bounded, we first show that $\hat{\delta}_t(\tilde{\pi}_t;\lambda)$ is uniformly bounded. Without loss of generality, consider $\hat{\delta}_{1t}(\tilde{\pi}_t;\lambda)$:

$$\begin{aligned}
\hat{\delta}_{1t}(\tilde{\pi}_t;\lambda) &= -\log \int \frac{\exp(w_{jt}'v)}{1+\sum_{k=1}^{J_t}\exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)+w_{kt}'v)\tilde{\pi}_{kt}} dF(v;\lambda) \\
&\geq -\log \int \exp(w_{jt}'v) dF(v;\lambda) \\
&\geq -\log \sup_{\lambda\in\Lambda} \sup_{\|w\|\leq \overline{w}} \int \exp(w'v) dF(v;\lambda),
\end{aligned}$$

where the second inequality holds by condition (i). Then by condition (ii), we have $\inf_{t,\lambda,\tilde{\pi}_t} \hat{\delta}_{1t}(\tilde{\pi}_t;\lambda) > -\infty$. To show that $\sup_{t,\lambda,\tilde{\pi}_t} \hat{\delta}_{1t}(\tilde{\pi}_t;\lambda) < \infty$, consider the outside share:

$$\tilde{\pi}_{0t} := 1 - \mathbf{1}_{J_t}'\tilde{\pi}_t = \int \frac{1}{1+\sum_{k=1}^{J_t}\exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)+w_{kt}'v)\tilde{\pi}_{kt}} dF(v;\lambda).$$

By $\|\tilde{\pi}_t - \pi_t\| < e_2 < \underline{\varepsilon}_0/(2\bar{J})$ and $\pi_{0t} \geq \underline{\varepsilon}_0$, we have $\tilde{\pi}_{0t} \geq \underline{\varepsilon}_0/2$. Then there must exists $\overline{v}$ large enough such that $\int_{\|v\|\leq\overline{v}} \frac{1}{1+\sum_{k=1}^{J_t}\exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)+w_{kt}'v)\tilde{\pi}_{kt}} dF(v;\lambda) \geq \underline{\varepsilon}_0/4$. Then

$$\begin{aligned}
\hat{\delta}_{1t}(\tilde{\pi};\lambda) &\leq -\log \int_{\|v\|\leq\overline{v}} \frac{\exp(w_{jt}'v)}{1+\sum_{k=1}^{J_t}\exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)+w_{kt}'v)\tilde{\pi}_{kt}} dF(v;\lambda) \\
&\leq -\log \left\{ \left[\min_{\|w\|\leq\overline{w},\|v\|\leq\overline{v}} \exp(w'v) \right] \int_{\|v\|\leq\overline{v}} \frac{1}{1+\sum_{k=1}^{J_t}\exp(\hat{\delta}_{kt}(\tilde{\pi}_t;\lambda)+w_{kt}'v)\tilde{\pi}_{kt}} dF(v;\lambda) \right\} \\
&\leq -\left[\min_{\|w\|\leq\overline{w},\|v\|\leq\overline{v}} (w'v) \right] - \log(\underline{\varepsilon}_0/4).
\end{aligned}$$

Thus, $\sup_{t,\lambda,\tilde{\pi}_t} \hat{\delta}_{1t}(\tilde{\pi}_t;\lambda) < \infty$.

Now we show that $b_t(\tilde{\pi}_t;\lambda)$ is uniformly bounded. Without loss of generality, consider the first element of $b_t(\tilde{\pi}_t;\lambda)$:

$$\tilde{\pi}_{1t}^{-2} \int \tilde{\pi}_{1t}(v)^2 dF(v;\lambda) = \int \left(\frac{\exp(w_{1t}'v + \hat{\delta}_{1t}(\tilde{\pi}_t;\lambda))}{1+\sum_{k=1}^{J_t}\exp(w_{kt}'v + \hat{\delta}_{kt}(\tilde{\pi}_t;\lambda))\pi_{kt}} \right)^2 dF(v;\lambda)$$

$$\leq \exp(2\hat{\delta}_{1t}(\tilde{\pi}_t;\lambda)) \int \exp(2w_{1t}'v) dF(v;\lambda).$$

Then by condition (ii) and $\sup_{t,\lambda,\tilde{\pi}_t} \hat{\delta}_{1t}(\tilde{\pi}_t;\lambda) < \infty$, we have

$$\sup_t \sup_\lambda \sup_{\|\tilde{\pi}_t - \pi_t\| \leq e_2} \|\tilde{\pi}_{1t}^{-2} \int \tilde{\pi}_{1t}(v)^2 dF(v;\lambda)\| < \infty.$$

Analogous arguments establish the uniform boundedness of the other elements of $b_t(\tilde{\pi}_t;\lambda)$. This concludes the proof. □

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Anderson, C. (2006): *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequality Models," *Econometrica*, 81.

Berry, S. (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, pp. 242–262.

Berry, S., A. Gandhi, and P. Haile (2013): "Connected substitutes and invertibility of demand," *Econometrica*, 81(5), 2087–2111.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, pp. 841–890.

Berry, S., J. Levinsohn, and A. Pakes (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Vehicle Market," *Journal of Political Economy*, 112, 68–104.

Berry, S., O. Linton, and A. Pakes (2004): "Limit theorems for estimating the parameters of differentiated product demand systems," *Review of Economic Studies*, 71(3), 613–654.

Berry, S., and A. Pakes (2007): "THE PURE CHARACTERISTICS DEMAND MODEL," *International Economic Review*, 48(4), 1193–1225.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Chevalier, J. A., A. K. Kashyap, and P. E. Rossi (2003): "Why Don't Prices Rise During Periods of Peak Demand? Evidence from Scanner Data," *American Economic Review*, 93(1), 15–37.

Dezhbakhsh, H., P. H. Rubin, and J. M. Shepherd (2003): "Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data," *American Law and Economics Review*, 5(2), 344–376.

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.

Freyberger, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185(1), 162–181.

Gabaix, X. (1999): "Zipf's Law and the Growth of Cities," *The American Economic Review, Papers and Proceedings*, 89, 129–132.

Gandhi, A., Z. Lu, and X. Shi (2013): "Estimating Demand for Differentiated Products with Error in Market Shares," *CeMMAP working paper*.

Goolsbee, A., and A. Petrin (2004): "The consumer gains from direct broadcast satellites and the competition with cable TV," *Econometrica*, 72(2), 351–381.

Head, K., T. Mayer, et al. (2013): "Gravity equations: Workhorse, toolkit, and cookbook," *Handbook of international economics*, 4.

Hosken, D., and D. Reiffen (2004): "Patterns of retail price variation," *RAND Journal of Economics*, pp. 128–146.

Kahn, S., and E. Tamer (2009): "Inference on Randomly Censored Regression Models Using Conditional Moment Inequalities," *Journal of Econometrics*, 152, 104–119.

Nevo, A., and K. Hatzitaskos (2006): "Why does the average price paid fall during high demand periods?," Discussion paper, CSIO working paper.

Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

Nurski, L., and F. Verboven (2016): "Exclusive Dealing as a Barrier to Entry? Evidence from Automobiles," *The Review of Economic Studies*, 83(3), 1156.

Pollard, D. (1990): "Empirical Process Theory and Application, NSF-CBMS Regional Conference Series in Probability and Statistics," II, Institute of Mathematical Statistics.

Quan, T. W., and K. R. Williams (2015): "Product Variety, Across-market Demand Heterogeneity, And The Value Of Online Retail," *Working Paper*.

Sherman, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.