University of Toronto
Department of Economics

Working Paper 801

Efficient Estimation of Structural Models via Sieves

By Yao Luo and Peijun Sang

June 22, 2025

Efficient Estimation of Structural Models via Sieves*

Yao Luo† Peijun Sang‡

†University of Toronto

‡University of Waterloo

June 14, 2025

Abstract

We propose a class of sieve-based efficient estimators for structural models (SEES), which approximate the solution using a linear combination of basis functions and impose equilibrium conditions as a penalty to determine the best-fitting coefficients. Our estimators circumvent repeated solution of the structural model, apply to a broad class of models, and are consistent, asymptotically normal, and asymptotically efficient. Moreover, they solve *unconstrained* optimization problems with fewer unknowns and offer convenient standard error calculations. As an illustration, we apply our method to an entry game between Walmart and Kmart.

Keywords: Efficient Estimation, Sieves, Empirical Games, Joint Algorithm, Nested Algorithm

*First version: February 2022. *Contact Information*: Luo: Department of Economics, University of Toronto, Max Gluskin House, 150 St. George St, Toronto, ON M5S 3G7, Canada (email: yao.luo@utoronto.ca); Sang: Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada, N2L 3G1, Canada (email: psang@uwaterloo.ca). We are grateful to Victor Aguirregabiria, Xiaohong Chen, Paul Grieco, John Rust, Marc Rysman and seminar participants at the Canadian Economic Association, ICSA-Canada, Asian Meeting of the Econometric Society, CESG 2023, University of Toronto and Western University for helpful discussions and comments. Luo acknowledges the Social Sciences and Humanities Research Council of Canada for research support. All errors are our own.

1 Introduction

A structural model builds on economic theory and describes how a set of endogenous variables are related to a set of explanatory variables. This relationship is typically characterized by an implicit function. In particular, it generates endogenous function p determined by an equation system:

$$p = \Psi(p, \theta), \tag{1}$$

where θ is the parameter of interest and Ψ is a representation of the structural model.[1] While Ψ is explicit, solving for p could be difficult or costly. Such computational burden limits the use of standard estimators. For instance, the maximum likelihood estimator (MLE) repeatedly guesses θ and evaluates data likelihood using the solution of (1), $p^*(\theta)$. However, finding the solution can be computationally intensive and is often hindered by a lack of robust algorithms, particularly in empirical games.

We introduce a new class of estimators: sieve-based efficient estimators for structural models (SEES). Our approach applies to a broad class of models, including empirical games, and avoids solving the model. Our SEES is motivated by two popular approaches to infinite-dimensional optimization problems: approximation and penalization. See Shen (1997), Shen (1998), and Chen (2007). Because the likelihood function $\ell(p^*, \theta; \text{data})$ involves an unknown function p^*, maximizing the likelihood with respect to p^* and θ may lead to an asymptotically inefficient estimator for the parameter, and the resulting estimator may not necessarily be close to the solution of (1). To address these issues, prior studies utilize sieves that are less complex but dense to approximate the original function space, and regularization that assumes smoothness of this function.

In this paper, we estimate structural models by approximating the solution to avoid solving the model and regularizing with the equilibrium conditions that are built into the model itself. By combining the data fitting and model fitting criteria,

[1] In discrete choice models, the parameter captures consumer preferences and the observable is consumer choice; in auctions, the parameter captures the value distribution and the observable is the bid distribution; in dynamic models, the parameter describes the agent's intertemporal tradeoff and the observable is intertemporal choice.

we formulate our penalized log-likelihood criterion,

$$\underbrace{\ell(\beta, \theta; \text{data})}_{\text{data likelihood}} - \underbrace{\omega \times \rho(\beta, \theta)}_{\text{penalization}},$$

where ℓ and ρ measure the data fitting and model fitting, respectively.[2] Moreover, $\beta \in \mathbb{R}^K$ and $\omega \in \mathbb{R}_+$ govern the approximation and the weighting, respectively. Instead of imposing stronger smoothness assumptions than typically implied by theory, our approach relies solely on the model to regularize the sieve approximation. The smoothing parameter ω explicitly captures the weighting of the data likelihood and the equilibrium condition, and the dimension of the approximation parameter β, denoted by K, balances computational cost and solution accuracy.

Allowing these tuning parameters to diverge at appropriate rates, the proposed parameter estimator of θ is consistent, asymptotically normal, and asymptotically efficient. Intuitively, by gradually updating the smoothing parameter, we shift the weight from the data to the equilibrium condition. At the minimum, a preliminary nonparametric estimate of p (by letting $\omega = 0$) constitutes a good starting value but is subject to issues with nonparametric estimates. When the smoothing parameter increases, more weight is given to the equilibrium condition. As model restrictions are more strongly enforced, the estimator converges to the MLE.

We prescribe several algorithms to implement SEES. The first is a *joint* algorithm that finds the combination of sieve approximations and model parameters that best explains the data and satisfies the equilibrium conditions. That is, we maximize the penalized log-likelihood function with respect to (β, θ). The second is a *nested* algorithm that consists of two main parts. First, for each model parameter θ, we find the sieve approximation of p that best explains the data and satisfies the equilibrium conditions. Second, based on the approximated solution, we find the model parameter that best fits the data. While the joint algorithm is attractive because it results in a single-level optimization problem, the nested algorithm is quite intuitive, resembling MLE.

Our estimator allows for discrete and continuous state/heterogeneity in the model to be estimated. The standard practice of discretizing continuous state variables or covariates leads to efficiency loss. Under mild regularity conditions, we show that

[2]While our idea extends to other types of estimators, we focus on likelihood-based ones here.

our estimator has the same asymptotic distribution as MLE in both cases. To our knowledge, we are the first to combine approximation and penalization in estimating structural models. While some studies have adopted approximation approaches, none combines them with penalization. Another important advantage of our method is that it produces standard errors in the same way as the standard MLE using the Fisher information matrix, which is of considerable convenience in empirical work. As a side product, we also derive a similar approach for the mathematical program with equilibrium constraints (MPEC) estimators, which provides a faster alternative than the bootstrapping method previously proposed by Su and Judd (2012).

We acknowledge several limitations inherent in our methodology. First, our sieve-based approach presupposes the smoothness of the solution within the state variables or covariates, leaving the treatment of discontinuities as a subject for future investigation. Second, our approach provides a robust solution that works with minimal assumptions on the solution, which is particularly valuable in models with unfavorable or unknown properties. However, it may not always be the most expedient choice in scenarios where the solution exhibits favorable properties, such as contraction mappings. For instance, as demonstrated in the simple model outlined in Section 2, it exhibits a relatively slower performance when compared to a nested-fixed point algorithm. Throughout this paper, we refrain from comparing computational time across different estimators, as it is often model-specific and, hence, more relevant in richer empirical models.

The remainder of the paper is organized as follows. Section 2 explains the idea using a simple example. Section 3 proposes the class of sieve-based efficient estimators for structural models and derives its asymptotic properties for the nested estimator in continuous states. Section 4 demonstrates the performance of our estimators in estimating an empirical game. Section 5 concludes. The Appendix contains all omitted proofs and details.

2 A Motivating Example

Our new method differs from existing methods by how we leverage data and model restrictions. We now compare it with popular existing methods, such as maximum likelihood estimation (MLE), two-step approaches, and nested pseudo-likelihood (NPL), through a motivating example.

Consider a monopolist j, facing logit demand, sells a product at a price P_j. That is, consumer i gets utility of

$$u_{ij} = \xi_j - \alpha P_j + \epsilon_{ij},$$

where ξ_j is *continuous* product quality, α is the price coefficient, and ϵ_{ij} represents the standard Type 1 extreme value (T1EV) taste shock. The firm's profit maximization problem is

$$\max_{P_j} \quad (\underbrace{P_j - c_j}_{\text{profit margin}}) \times \underbrace{\frac{\exp(\xi_j - \alpha P_j)}{1 + \exp(\xi_j - \alpha P_j)}}_{\text{market share}},$$

where c_j represents the constant marginal cost. The optimal price is determined by the FOC,

$$\alpha(P_j - c_j) = 1 + \exp(\xi_j - \alpha P_j),$$

where P_j appears both inside and outside an exponential function. As a result, the mapping from the parameters to the optimal price is implicit.

2.1 Structural Estimation

For simplicity, we focus on estimating the parameter θ that governs consumer preferences over product feature $x_j \in \mathbb{R}$ using observed prices. Specifically, we treat it as known that $c_j = 0,\ \alpha = 1,\ \xi_j = \log x_j + \log \theta + 1$, where "1" is quality normalization for simplicity. Appendix A shows that the optimal price satisfies

$$y_j^* = p(x_j; \theta), \tag{2}$$

where $y_j^* = P_j^* - 1$ represents the normalized price and $p(x_j; \theta)$ is defined by

$$p(x_j;\theta)e^{p(x_j;\theta)} = \theta x_j \quad \text{or} \quad p(x_j;\theta) = \theta x_j e^{-p(x_j;\theta)},$$

the first of which has the standard form of the Lambert W function[3] and the second of which has the same form as Equation (1). We denote this function as $p(\cdot;\theta)$ to indicate its dependence on the parameter.

Consider a data generating process (DGP) that is a noisy measurement of the

[3]The Lambert W function $W(x)$ is defined by $W(x)e^{W(x)} = x$.

optimal price $y_j = y_j^* + e_j$, where e_j's are measurement errors that are i.i.d. draws from the standard normal distribution. Therefore, the observed (normalized) price y_j is from the standard normal distribution with a location $p(x_j; \theta_0)$:

$$y_j \sim \mathcal{N}(p(x_j; \theta_0), 1), \text{ where } j = 1, \ldots, n.$$

The data contain the product characteristics x_j and the prices $P_j = P_j^* + e_j$ (equivalently, the normalized prices y_j). The parameter of interest is θ.

Maximum Likelihood Estimation: The standard MLE solves the following problem

$$\max_{\theta} \quad \sum_{j=1}^{n} \log \phi(y_j - p(x_j; \theta)),$$

where $\phi(\cdot)$ represents the density function of the standard normal distribution. Because $p(\cdot; \theta)$ is implicitly defined, this estimator is computationally costly. For each trial of θ, we need to find $p(x_j; \theta)$ for each data point x_j. The number of equations that need to be solved equals the sample size multiplied by the number of likelihood function evaluations.

Despite its asymptotic efficiency, the standard MLE requires solving the model for each parameter and thus solution algorithms that are sufficiently efficient and robust. When a contraction mapping solution for the model is available, it is often referred to as the nested fixed point algorithm (NFXP). See, e.g., Rust (1987) in dynamic discrete choice models and Berry et al. (1995) in demand models.

A Two-Step Approach: We can "invert the FOC" and obtain a representation of the "unknown" in terms of the optimal prices:

$$\theta = \frac{p(x_j; \theta) e^{p(x_j; \theta)}}{x_j},$$

where the normalized price $y_j^* = p(x_j; \theta)$ is unobserved. In principle, this FOC inversion allows estimating the parameter using the optimal price in any market.

Due to measurement errors, the rewritten FOC suggests a simple two-step approach that avoids solving the model repeatedly in estimation. In the first step, we consider $y_j^* = p(x_j; \theta_0)$ and estimate the optimal price as a function of the covariate.

Although the true parameter θ_0, the endogenous variable p and thus the RHS $p(x_j;\theta_0)$ are all unobserved, we can estimate the LHS $y_j^*(x_j)$ nonparametrically using the observed price and covariate pairs $\{x_j, y_j\}_{j=1}^n$. In particular, we run a nonparametric regression[4],

$$y_j = y_j^*(x_j) + e,$$

and obtain an estimate of the normalized price $\widehat{y_j^*(x_j)}$. In the second step, we have a simple plug-in estimator,

$$\widehat{\theta} = \text{median}\left\{ \frac{\widehat{y_j^*(x_j)} \times \exp \widehat{y_j^*(x_j)}}{x_j} : j = 1, \ldots, n \right\}.$$

Two-step approaches avoid repeatedly solving the economic model at the expense of efficiency. In the first step, the analyst obtains a nonparametric estimate of the endogenous variable $\widehat{p}$. In the second step, the estimate is obtained from $\widehat{p} = \Psi(\widehat{p}, \theta)$ in various ways. In auction models, Guerre et al. (2000) use the estimated bid distribution to construct pseudo values, which are then used to estimate the underlying value distribution. In dynamic discrete choice models, the conditional choice probability (CCP) approach of Hotz and Miller (1993) plugs the estimated CCPs into the optimal decision rules. In dynamic games, one can obtain a nonlinear least squares estimate of θ by replacing p with the estimated CCP $\widehat{p}$ in the function; see Pesendorfer and Schmidt-Dengler (2008).

Nested Pseudo-Likelihood Algorithm: In each iteration, the NPL algorithm solves the following problem:

$$\max_{\theta} \quad \sum_{j=1}^{n} \log \phi(y_j - \theta x_j \exp(-\widehat{p}_j^{\tau})),$$

where $\widehat{p}_j^{\tau}$ represents some estimate of the optimal price in market j. Denote the solution as $\widehat{\theta}^{\tau}$. We can then update the price estimates $\widehat{p}_j^{\tau+1} = \widehat{\theta}^{\tau} x_j \exp(-\widehat{p}_j^{\tau})$. We iterate the process till the parameter estimate converges.

Given some estimates $\widehat{\theta}$ and $\widehat{p}$, the NPL algorithm obtains new estimates of the choice probabilities by applying the mapping $\tilde{p} = \Psi(\widehat{p}, \widehat{\theta})$ and then updates the pa-

[4]We apply kernel regression using the optimal bandwidth estimated by cross-validation. See, e.g., Hall et al. (2004), Li and Racine (2004) and Hall and Racine (2015).

rameter estimate by maximizing the pseudo-likelihood function $\ell(\tilde{p}, \theta)$.

A Sieve-Based Efficient Estimator: In this paper, we propose a method that obviates solving the model repeatedly. In particular, we approximate the "solution" function $p^\star(\cdot)$ by B-spline basis functions:

$$p^\beta(\cdot) = \sum_{k=1}^{K} \beta_k s_k(\cdot),$$

where $s_k(\cdot)$ is a cubic spline basis function, and K denotes the number of basis functions. Our sieve-based estimator of θ maximizes the likelihood

$$\sum_{j=1}^{n} \log \left\{ \phi \left(y_j - p^{\widehat{\beta}(\theta,\omega)}(x_j) \right) \right\},$$

where $\widehat{\beta}(\theta, \omega)$ is defined by

$$\arg\max_{\beta \in \mathbb{R}^K} \quad \sum_{j=1}^{n} \log \left\{ \phi(y_j - p^\beta(x_j)) \right\} - \underbrace{\omega \int_{\mathcal{X}} \left[p^\beta(x) e^{p^\beta(x)} - \theta x \right]^2 dx}_{\text{penalization}},$$

where $\omega > 0$. Because $p^\beta(\cdot)$ is an approximation, the second term penalizes the likelihood by the amount of deviation by definition of the Lambert W function. Importantly, this penalization does not depend on the observed data; it is driven solely by how much the approximation violates the equilibrium conditions.

Discussion

We now compare the above-mentioned estimators. First, SEES and MLE are asymptotically equivalent and almost identical in finite samples. However, NFXP algorithms may converge slowly, and such mappings may not even exist in important models. For instance, empirical games, such as asymmetric auctions and dynamic games, are notoriously difficult to solve, making MLE difficult to apply. In contrast, we avoid solving the model repeatedly by approximating the solution flexibly.

Second, two-step approaches are limited by the first-step nonparametric estimation of the endogenous variable and may suffer from the "curse of dimensionality" when x has multiple dimensions (Stone, 1980). As a result, the finite-sample esti-

mation error can be substantial. In contrast, our approximated solution avoids this issue, as its final version relies almost entirely on the model.

Third, our estimator is also related to the nested pseudo-likelihood algorithm proposed by Aguirregabiria and Mira (2002, 2007). Exploiting the unique feature of dynamic discrete choice models that the Jacobian matrix of Ψ_θ is always zero, their iterative refinement converges to MLE. However, it requires some discretion in applying it to empirical games. See Pesendorfer and Schmidt-Dengler (2010). Both algorithms bridge the gap between the standard MLE and two-step methods, and are asymptotically equivalent to MLE. However, they are based on very different ideas. Our estimator is flexible to accommodate different estimation algorithms, including one that resembles NPL, and robust in applications to various models, including empirical games.

To our knowledge, we are the first to combine approximation and penalization in estimating structural models. While some studies have adopted approximation approaches, none combines them with penalization. For instance, Keane and Wolpin (1994, 1997) use sieves to approximate solutions in dynamic structural models, combining approximation and NFXP. In estimating dynamic games, Sweeting (2013) uses parametric approximations to the value function, combining approximation and NPL. Most related, Barwick and Pathak (2015) approximates the value function using sieves and imposes the Bellman equation as an equilibrium constraint.

Another related algorithm is MPEC, which is an alternative computational algorithm to MLE. See, e.g., Su and Judd (2012) and Dubé et al. (2012).[5] It avoids solving the model repeatedly by augmenting the unknown to (θ, p) and imposing the equilibrium condition as a constraint:

$$\begin{aligned} \max_{\theta,p} \quad & \ell(p, \theta; \text{data}) \\ \text{subject to} \quad & p = \Psi(p, \theta). \end{aligned}$$

We will show that our SEES's dual problem is a natural extension of MPEC in discrete state settings. Our SEES nests MPEC as a limiting case when the number of basis functions is the same as the dimension of p and the regularization parameter

[5]Several papers have compared MPEC with the original estimators for various models. For a comparison of NFXP and MPEC, see, e.g., Lee and Seo (2015) for the Berry et al. (1995) model and Iskhakov et al. (2016) for dynamic models.

equals infinity. MPEC forms the Lagrangian function using Lagrange multipliers Λ that are of the same size as p: $\max_{\theta,p,\Lambda} \ell(p,\theta;\text{data}) - \Lambda'(p - \Psi(p,\theta))$. As a result, it solves $2 \times \dim(p) + \dim(\theta)$ equations in the same number of unknowns. Our SEES approximates p by β and introduces a scalar regularization parameter ω, which reduces the problem to an *unconstrained* optimization problem with $K + 1 + \dim(\theta)$ unknowns. Therefore, SEES is computationally convenient because $K + 1 \ll 2 \times \dim(p)$.

2.2 Simulation Evidence

Consider $x_j \sim \text{Uniform}[0, \overline{x}]$ and $\theta_0 = 1$. For MLE, we use the bisection method to solve for the Lambert W function. We omit MPEC here for two reasons. First, we focus on the statistical properties rather than the computational time of the estimators. Second, MPEC is an alternative computational algorithm to MLE with identical statistical properties. For the two-step approach, we use local linear kernel regression and apply the optimal bandwidth chosen via cross validation. For the proposed method, we use the cubic spline explained in Luo et al. (2018) and let $K = 6$; the choice of ω follows the method that we propose later. We also provide the analytic gradient of the outer loop and the analytic gradient and Hessian of the inner loop maximization problem; see Appendix F.1.

Table 1 shows the simulation results of 1,000 replications with a sample size of 1,000. SEES, MLE, and NPL perform very similarly. In particular, SEES and MLE are almost identical in each replication. Figure 1 compares MLE and the iterations of NPL and SEES in a typical replication.[6] While their earlier iterations could differ from MLE significantly, NPL and SEES both converge to a close neighbourhood of MLE.

In contrast, the two-step approach generates a larger bias and standard error. Alternatively, we can take the sample average in the second step. However, the noise in nonparametric estimates near boundaries deteriorates the estimates substantially. The median performs much better the mean. It is clear that the performance of the two-step estimator is affected by the first-stage nonparametric regression.

Remark 1. This simple model has a convenient property that each firm's optimal

[6]SEES usually converges in 2-4 iterations using our proposed choice of ω. For better visualization in this figure, we increase $\log \omega$ in 7 equal steps to match the number of iterations of NPL.

Figure 1: Compare MLE, NPL, and SEES

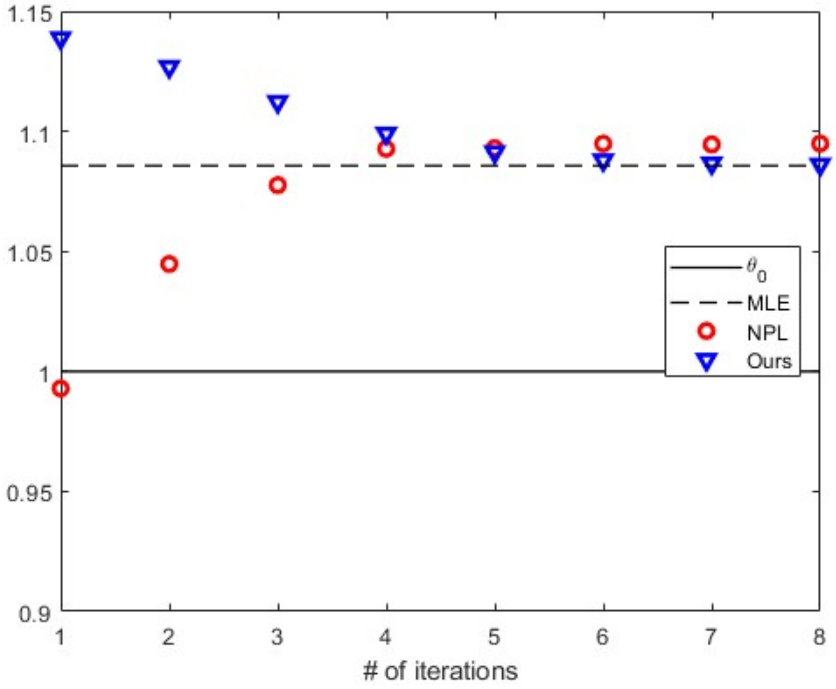


Table 1: Monopoly Pricing: $\overline{x} = 1$

	SEES	MLE	NPL	2-Step	
				median	mean
mean	1.0029	1.0030	1.0030	0.9639	1.4179
se	0.1282	0.1283	0.1285	0.1356	2.8309

price is the solution of a strictly monotone function. In this case, the bisection method is fast and robust in solving the model, except that it takes many iterations to converge.[7] In empirical games, the equilibrium is the solution of a system of nonlinear equations, which is harder to find and often lacks reliable algorithms. Appendix E provides additional simulations using a much richer DGP motivated by our empirical application of static games.

[7]Alternatively, we show that the Lambert W function $W(x)$ can be calculated using the contraction mapping $\Psi(W, x) = xe^{-W}$ when x is smaller than Euler's number.

3 Our Sieve-Based Efficient Estimator

We now describe our estimator in detail. Let $x \in \mathbb{R}^I$ denote the state or heterogeneity, and let $p^*(x;\theta) \in \mathbb{R}^J$ denote the endogenous solution, where I and J are positive integers. Given any $\theta \in \Theta \subset \mathbb{R}^d$, $p^*(\cdot;\theta) \in \mathbb{R}^J$ denotes the solution to the following the structure equation:

$$p(\cdot) = \Psi(p(\cdot), \theta), \tag{3}$$

Rather than solving for $p^*(\cdot;\theta)$ at each value of θ during likelihood evaluation, we approximate the true solution by p^β. The choice of approximation infrastructure depends on its approximation properties and computational convenience. A popular one often adopted in empirical studies is the method of sieves. When x is univariate, we can use a standard series expansion $p^\beta(x) = \sum_{k=1}^K \beta_k s_k(x)$, where $\{s_1, \ldots, s_K\}$ are basis functions spanning the finite-dimensional sieve space $\mathcal{B}$. When x is multivariate, we can use a tensor-product sieve $p^\beta(x) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_I=1}^{K_I} \beta_{k_1,\ldots,k_I}\, s_{k_1}(x_1) \cdots s_{k_I}(x_I)$. For convenience, we denote $K = (K_1, \ldots, K_I)'$ and refer to it as the approximation parameter. Typical choices of s_k include B-spline basis functions and Bernstein polynomials. Such methods are flexible in accounting for shape restrictions imposed by the structural model, such as nonnegativity and monotonicity. For instance, if p represents choice probabilities, we can use $p^\beta(\cdot) = [1 + \exp\{\sum_{k=1}^K \beta_k s_k(\cdot)\}]^{-1}$ to ensure that $p^\beta \in (0,1)$.

Moreover, our SEES imposes the model constraints by penalizing the difference between p^β and $\Psi(p^\beta, \theta)$. This difference is independent of the data sample in measuring the fidelity of approximation to the equilibrium conditions. The smaller the difference, the better the approximate p^β satisfies equilibrium conditions.

We formulate the penalized log-likelihood criterion by combining the data fitting and model fitting criteria,

$$\underbrace{\ell\big[p^\beta(\cdot), \theta\big]}_{\ell(\beta,\theta)} - \omega \times \underbrace{\rho\Big[p^\beta(\cdot), \Psi\big(p^\beta(\cdot), \theta\big)\Big]}_{\rho(\beta,\theta)}, \tag{4}$$

where $\omega > 0$ is the smoothing parameter, and ρ is a metric that measures the difference between p^β and $\Psi(p^\beta, \theta)$. For instance, when $p^\beta(x)$ is scalar, we can use the Euclidean norm $\rho\big(p, \Psi(p,\theta)\big) = \|p - \Psi(p,\theta)\|_2^2$. If instead $p^\beta(x) = \big(p_1^\beta(x), \ldots, p_J^\beta(x)\big)$ is J-

dimensional, we can use $\rho\big(p, \Psi(p,\theta)\big) = \sum_{j=1}^{J} \|p_j - \Psi_j(p,\theta)\|_2^2$.[8]

3.1 Estimation Algorithms

We develop two algorithms to implement our estimator given each smoothing parameter ω: a joint algorithm and a nested algorithm.

Joint Algorithm This algorithm is attractive because it involves a single-level optimization problem. We augment the unknown to (β, θ) and solve the following problem:

$$\max_{\beta,\theta} \quad \ell(\beta,\theta) - \omega \times \rho(\beta,\theta), \tag{5}$$

which leads to $\widehat{\beta}(\omega)$ and $\widehat{\theta}(\omega)$. We recommend providing analytical gradients and Hessians to improve computational efficiency and estimation accuracy.

Nested Algorithm This algorithm is intuitive, resembling MLE. There are two layers of optimization problems to be solved. In the inner layer, given (θ, ω), we find the best approximation parameter $\widehat{\beta}(\theta;\omega)$ that solves the following problem:

$$\max_{\beta\in\mathbb{R}^K} \quad \ell(\beta,\theta) - \omega \times \rho(\beta,\theta). \tag{6}$$

Solving (6) indicates that the maximizer $\widehat{\beta}(\theta;\omega)$ is an implicit function of θ.

In the outer layer, applying the best fitting approximation parameter, we search for the structural parameter $\widehat{\theta}(\omega)$ that maximizes the following likelihood:

$$\ell\left(\widehat{\beta}(\theta,\omega),\theta\right). \tag{7}$$

Note that we have considered the structural equation (1) in the inner layer. Therefore, the equilibrium conditions are embedded in $\widehat{\beta}(\theta,\omega)$. As illustrated above, the optimizer of (6), $\widehat{\beta}(\theta;\omega)$, is an implicit function of θ. Therefore, $\ell\left(\widehat{\beta}(\theta,\omega),\theta\right)$ in (7) is a function of θ. We obtain the final estimator of θ, denoted by $\widehat{\theta}(\omega)$, by directly maximizing (7) with respect to θ.

Remark 2. Because we solve the inner loop problem many times, it is more efficient

[8] Component-specific smoothing is easily accommodated by modifying the penalty to $\sum_{j=1}^{J} \omega_j \|p_j - \Psi_j(p,\theta)\|_2^2$.

to provide the gradient and Hessian of $h(\beta, \theta) = \ell(\beta, \theta) - \omega\rho(\beta, \theta)$ with respect to β, as well as the gradient of the objective function in the outer loop $\widehat{\ell}(\theta) = \ell(p^{\widehat{\beta}(\theta)}(\cdot), \theta)$ with respect to θ. While the former is often straightforward, the latter is a bit more involved because the best-fitting approximation $\widehat{\beta}(\theta)$ is implicit. In particular, it requires deriving how the best-fitting approximation changes with respect to the model parameter, $\nabla\widehat{\beta}(\theta)$. Proposition A1 in Appendix D provides the gradient of the outer loop.

Alternatively, we can use an *alternating iterative* algorithm in place of either algorithm. That is, we can iterate the problem in (6) given a current estimate of θ and the problem in (7) to update the estimate of θ. The iterations will continue until convergence. This iterative approach is similar to NPL.[9]

The Choice of Tuning Parameters

We propose a new method that selects the smoothing parameter ω based on the performance of parameter estimation. Intuitively, we choose a sufficiently large ω to ensure fidelity in approximating the equilibrium conditions.[10] In particular, we start with a moderate ω_1 and update it till the estimates converge. For each $\omega = \omega_\tau$, we can conduct the joint or nested algorithm and obtain an estimate $\widehat{\theta}(\omega_\tau)$. Consider a significance level of α. We obtain its standard error $\widehat{\sigma}(\omega_\tau)$ and confidence interval $\mathcal{I}(\omega_\tau) = [\widehat{\theta}(\omega_\tau) + z_{\alpha/2}\widehat{\sigma}(\omega_\tau), \widehat{\theta}(\omega_\tau) + z_{1-\alpha/2}\widehat{\sigma}(\omega_\tau)]$, applying the standard formula of standard error calculation for MLE.

We multiply the smoothing parameter by L each time, i.e. $\omega_{\tau+1} = L \times \omega_\tau$ and obtain a new estimate and its confidence interval $\mathcal{I}(\omega_{\tau+1})$. Continue this process till the overlapping portion of the two intervals accounts for more than a threshold percentage of both of the two. That is, our final choice of the smoothing parameter is $\widehat{\omega} = \omega_\tau$ if

$$\min\left\{\frac{|\mathcal{I}(\omega_\tau) \cap \mathcal{I}(\omega_{\tau-1})|}{|\mathcal{I}(\omega_{\tau-1})|}, \frac{|\mathcal{I}(\omega_\tau) \cap \mathcal{I}(\omega_{\tau-1})|}{|\mathcal{I}(\omega_\tau)|}\right\} \geq \mathfrak{c},$$

where $|\cdot|$ represents the length of the interval. In case the parameter of interest θ is

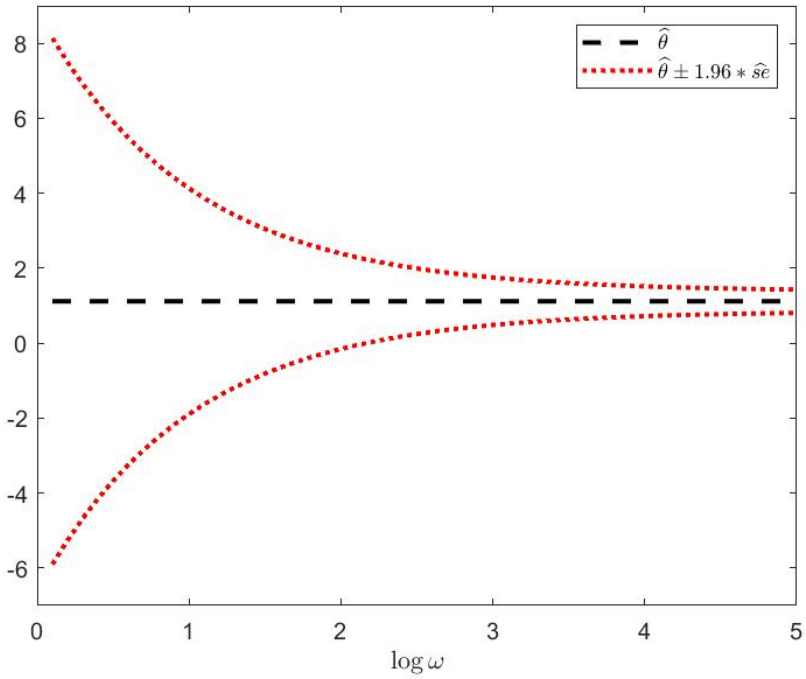
[9] Our algorithms, like other methods, do not guarantee finding global optima. We recommend experimenting with different starting points.

[10] Cross-validation is commonly used to balance bias and variance in estimation or prediction when the function of interest is unknown, by estimating prediction error or evaluating the likelihood function on held-out data. However, in our case, this trade-off is not a concern because the function $p(\cdot)$ is fully specified by the structural model (1).

multi-dimensional, we check this condition element by element. The final estimate of the model parameter follows $\widehat{\theta}(\widehat{\omega})$.

All the simulations reported in this paper adopt $\alpha = 0.05$, $L = 10$, and $\mathfrak{c} = 95\%$. Therefore, $z_{\alpha/2} = -1.96$ and $z_{1-\alpha/2} = 1.96$. To illustrate how the proposed method works in terms of selecting the tuning parameter, consider the monopoly pricing example with $x_j = 1$. Figure 2 reports the parameter estimate and confidence intervals when the smoothing parameter varies. The x-axis represents $\log \omega$. While the bias seems small for small values of ω, the confidence interval is large. As ω increases, it shrinks to the MLE confidence interval.

Figure 2: Choosing The Smoothing Parameter



Note: The DGP is $y_j \sim \mathcal{N}(W(\theta_0), 1)$, where $\theta_0 = 1$. This figure demonstrates how the estimate and its confidence interval change when the smoothing parameter increases.

The approximation parameter K (i.e., the number of basis functions) should be chosen properly: large enough to approximate the equilibrium well. Exactly how many is sufficient depends on the complexity of the equilibrium solution. In our motivating example, the solution is simple; as a result, we find that four cubic basis functions are adequate to approximate the solution well. When the solution is complex and the number of basis functions needs to be large, the analyst should start with a

large smoothing parameter to avoid over-fitting in the inner loop; i.e., the likelihood function dominates. Sometimes, there are a finite number of states in the structural model, which is often assumed in estimating dynamic models. See, e.g., Aguirregabiria and Mira (2002) and Pesendorfer and Schmidt-Dengler (2008). Such finite states often come from discretization of covariates. In this case, the approximation can be perfect, i.e., $\beta = p$. That is, $s_k(x) = \mathbb{1}(x = p_k)$, where $\mathbb{1}(\cdot)$ is the indicator function and p_k is the kth element in the endogenous variable p. Note in this scenario the number of basis functions is identical to the dimensionality of p.

To capture empirically relevant covariates without losing much efficiency, any approximation methods would suffer from a computational curse of dimensionality — the total number of basis functions has to grow fast as the dimensionality of x increases. We propose to resolve this issue in several ways. First, more advanced approximation methods are often preferable to simple ones. See, e.g., Chen et al. (2023a) compare neural networks-based estimators. Additional shape constraints, sparsity patterns, and better grid choices are useful in reducing computational burden. See, e.g., Chen (2007) discusses various sieve-based methods, and Kristensen et al. (2021) discuss various approximation architectures for approximating value functions in dynamic models. Second, there are also many model-specific techniques for approximating functions using a small number of basis functions. The model may generate multiple endogenous objects, some directly observable while others intermediate. A well-chosen p simplifies its approximation and evaluating Ψ and data likelihood. For instance, in static games of asymmetric information, if the deterministic component in the payoff function is linear in the parameters, see, e.g., Bresnahan and Reiss (1991) and Bajari et al. (2010), how covariates determine the endogenous variable becomes a multiple-index model. In empirical auctions, Chen et al. (2023b) approximate the bid-stage primitives by flexible Bernstein polynomial sieves. One can borrow techniques from the existing literature in estimating such a model.

3.2 Asymptotic Properties in Continuous State Settings

In this section, we study the asymptotic properties of the nested SEES algorithm when the states are continuous. Let x denote the continuous states. Without loss of generality, we assume that $I = 1$ and $x \in [0, T]$ with a fixed $T > 0$. Without loss of generality, we assume $T = 1$. We approximate the solution to (3) using a

sieve method. In particular, we take the sieve space, denoted as $\mathcal{B}_n$, to be the space of cubic B-spline functions equipped with knots $\tau^{(n)} = \left\{0 = t_0^{(n)} < \cdots < t_{M_n}^{(n)} = T\right\}$. Let $|\tau^{(n)}| = \max_{0 \le i \le M_n - 1} |t_{i+1}^{(n)} - t_i^{(n)}|$ be the largest distance of adjacent knots in $\tau^{(n)}$. For any element $\eta \in \mathcal{B}_n$, there exists a $\beta = (\beta_1, \ldots, \beta_K)^\top \in \mathbb{R}^K$ such that $\eta(x) = \sum_{j=1}^K \beta_j s_j(x)$, where s_j's are cubic B-spline basis functions and $K = M_n + 3$.

Suppose that $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d observations, where X_i's are independently sampled from a distribution Q on $[0, T]$ and Y_i's take values in $\mathbb{R}^{d_Y}$. For simplicity, we assume that $d_Y = 1$ and $J = 1$. Following the notations defined in the last subsection, given $\theta \in \Theta$, we have

$$\hat{p}_n(\cdot; \theta) = \arg\max_{p \in \mathcal{B}_n} \ell_n(p(\cdot)) - \omega \rho[p(\cdot), \Psi(p(\cdot), \theta)], \tag{8}$$

where the likelihood ℓ_n can be written

$$\ell_n(p(\cdot)) = \frac{1}{n} \sum_{i=1}^n f(Y_i, p(X_i)),$$

for any function p defined on $[0, T]$, and the penalty function ρ is given by

$$\rho[p(\cdot), \Psi(p(\cdot), \theta)] = \int_0^T \{p(x) - \Psi(p(x), \theta)\}^2 \, \mathrm{d}x.$$

Then the nested estimator of θ is given by

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\hat{p}_n(\cdot; \theta)). \tag{9}$$

We next study the asymptotic properties of $\hat{\theta}_n$ in this model. We first establish consistency for the estimator and then develop the asymptotic normality of $\hat{\theta}_n$. To this end, we need the following regularity conditions.

Assumption 1. The density function of Q, denoted by q, satisfies that $C_1 \le q(x) \le C_2$ for any $x \in [0, T]$, where C_1 and C_2 are two positive constants.

Assumption 1 ensures that X_i's are evenly distributed over $[0, T]$, which is entailed by a good estimation of p over the entire domain. Moreover, this assumption is commonly adopted in the literature of nonparametric smoothing; see Stone (1985) and Chen et al. (2023b) for example.

Assumption 2. The parameter space Θ is a compact subset of $\mathbb{R}^d$.

Define $g(p(x), \theta) = \Psi(p(x), \theta) - p(x)$ for any function p defined on $[0, T]$. Let $h^{(k)}$ denote the kth order derivative of function h for any integer $k \geq 0$, and define

$$C^k([0,T]) = \{h : h^{(k)} \text{ is continuous on } [0,T]\}.$$

Obviously, the solution to Equation (1), denoted as $p^*(\theta)$, satisfies $g(p^*(\theta), \theta) = 0$. We impose the following regularity condition on $g(p, \theta)$.

Assumption 3. For any $\theta \in \Theta$, $\Psi \in C^4(\mathbb{R} \times \Theta)$. There exists a positive constant r such that for any $(p_1, \theta), (p_2, \theta)$ satisfying $\max\{\|p_1\|_\infty, \|p_2\|_\infty\} \leq r$ and $\theta \in \Theta$, we have $\|p_1 - p_2\|_{L^2([0,T])} \leq C_g \|g(p_1, \theta) - g(p_2, \theta)\|_{L^2([0,T])}$ for some constant $C_g > 0$.

By Lemma A1 in Appendix B, we define the sieve space to be

$$\mathcal{B}_n(r) = \left\{ \eta(x) : \eta(x) = \sum_{k=1}^{K} \beta_k s_k(x), \|\eta\|_\infty \leq r \right\}$$

with equally spaced knots for some sufficiently large constant r. Therefore, for any $\theta \in \Theta$, there exits an $p_{\theta,n} \in \mathcal{B}_n(r)$ such that $\|p^*(\cdot; \theta) - p_{\theta,n}\|_\infty = O(K^{-4})$. Let

$$r_n = \sup_{\theta \in \Theta} \inf_{\eta \in \mathcal{B}_n(r)} \|p^*(\cdot; \theta) - \eta\|_\infty. \tag{10}$$

For the sieve estimator $\hat{p}_n(\cdot; \theta)$ defined in (8), we establish an important approximation error bound, which will be used to develop the asymptotic normality for $\hat{\theta}_n$ defined in (9) later.

Theorem 1. *Assume that Assumptions 1-3 and 5 are satisfied. Furthermore, if $\omega \to \infty$ as $n \to \infty$, we have*

$$\sup_{\theta \in \Theta} \|\hat{p}_n(\cdot; \theta) - p^*(\cdot; \theta)\|^2_{L^2(\mathbb{P})} := \sup_{\theta \in \Theta} \int_0^T \{\hat{p}_n(x; \theta) - p^*(x; \theta)\}^2 q(x) \, \mathrm{d}x = O_{\mathbb{P}}(\omega^{-1}) + O_{\mathbb{P}}(r_n^2).$$

Let θ_0 denote the true value of θ and G_{θ_0} denote the joint distribution of (X_i, Y_i) under this true value. Define

$$M(\theta) = \mathrm{E}_{\theta_0}[f(Y_i, p^*(X_i; \theta)], \qquad \theta \in \Theta,$$

where the expectation is taken with respect to G_{θ_0}.

Assumption 4. $M(\theta)$ is continuous function and has a unique maximum at θ_0 in Θ.

Assumption 5. $f(y,p) \leq 0$ for any $(y,p) \in \mathbb{R}^2$ and $f(y,p) \in C(\mathbb{R} \times \mathbb{R})$ satisfies

$$\mathrm{E}_{\theta_0}\left[\left|\frac{\partial f}{\partial p}(Y_i, p^*(X;\theta_0))\right|\right] < \infty.$$

Moreover, if Y_i is not bounded, $f(y,p)$ satisfies that for any compact set $D \subset \mathbb{R}$,

$$\liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}[-f(y,p)]}{\sup_{p\in D}[-f(y,p)]} > 0 \quad \text{and} \quad \liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}[-\partial f(y,p)/\partial p]}{\sup_{p\in D}[-\partial f(y,p)/\partial p]} > 0$$

To establish consistency and asymptotic normality for $\hat{\theta}_n$, a stronger version of Assumption 5 is entailed.

Assumption 6. $f(y,p) \leq 0$ for any $(y,p) \in \mathbb{R}^2$ and $f(y,p) \in C(\mathbb{R} \times \mathbb{R})$ satisfies that under G_{θ_0}, $\partial f(Y_i, p^*(X_i;\theta_0))/\partial p$ is sub-Gaussian, and

$$\mathrm{E}_{\theta_0}\left[\left|\frac{\partial^2 f}{\partial p^2}(Y, p^*(X;\theta_0))\right|\right] < \infty.$$

Moreover, if Y_i is not bounded, $f(y,p)$ satisfies that for any compact set $D \subset \mathbb{R}$,

$$\liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}[-f(y,p)]}{\sup_{p\in D}[-f(y,p)]} > 0 \qquad \liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}|\partial f(y,p)/\partial p|}{\sup_{p\in D}|\partial f(y,p)/\partial p|} > 0,$$

and

$$\liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}|\partial^2 f(y,p)/\partial p^2|}{\sup_{p\in D}|\partial^2 f(y,p)/\partial p^2|} > 0.$$

Remark 3. We impose a sub-Gaussian condition on $\partial f(Y_i, p^*(X_i;\theta_0))/\partial p$ in Assumption 6, which is stronger than that in Assumption 5. This term is usually referred to as the residual in the gradient descent algorithm. Actually, this condition is met when $\epsilon_i = Y_i - p(X_i)$ follows a normal distribution and $f(Y_i, p(X_i)) = -(Y_i - p(X_i))^2$ or in logistic regression when $\mathbb{P}(Y_i = 1|X_i) = \exp\{p(X_i)\}/[1+\exp\{p(X_i)\}]$. We impose this condition to ensure that $\ell_n(p^*(\cdot;\hat{\theta}_n)) \geq \sup_{\theta\in\Theta} \ell_n(p^*(\theta)) - o_{\mathbb{P}}(n^{-1})$. Alternatively, we may impose a stronger smoothness condition on $p^*(\cdot;\theta)$. Then a smaller approximation error, i.e., a smaller r_n (defined in (10)), can be obtained with a sieve space with higher-order B-spline functions. This reflects a trade-off between

the smoothness of the structure equation and the decaying rate of the tail probability of $\partial f(Y_i, p^*(X_i;\theta_0))/\partial p$.

Theorem 2. *Suppose Assumptions 1-4 and 6 hold. If* $\omega \to \infty$ *and* $K^2\log(K) = o(n)$, *then* $\hat{\theta}_n$ *is a consistent estimator of* θ_0.

Theorem 3. *Suppose that Assumptions 1-4 and Assumption 6 hold. If* $\omega/n^2 \to \infty, n^{1/4} = o(K)$, $K^2\log(K) = o(n)$ *and the matrix*

$$V_{\theta_0} = -\mathrm{E}_{\theta_0}\left[\frac{\partial f(Y_i, p^*(X_i;\theta_0))}{\partial p}\frac{\partial^2 p^*(X_i;\theta_0)}{\partial\theta\partial\theta'} + \frac{\partial^2 f(Y_i, p^*(X_i;\theta_0))}{\partial p^2}\left\{\frac{\partial p^*(X_i;\theta_0)}{\partial\theta}\right\}\left\{\frac{\partial p^*(X_i;\theta_0)}{\partial\theta}\right\}'\right]$$

is invertible, then

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

where $\Sigma = V_{\theta_0}^{-1}\mathrm{E}_{\theta_0}\left[\left\{\frac{\partial f(Y,p^*(X;\theta_0))}{\partial p}\right\}^2\left\{\frac{\partial p^*(X_i;\theta_0)}{\partial\theta}\right\}'\left\{\frac{\partial p^*(X_i;\theta_0)}{\partial\theta}\right\}\right]V_{\theta_0}^{-1}$.

Remark 4. 1) If the sieve space $\mathcal{B}_n(r)$ consists of the cubic spline functions equipped with equally spaced knots $\tau^{(n)}$, then by Lemma A1 in Appendix B, this condition implies $r_n = o(n^{-1})$. Meanwhile, we impose an upper bound on K to control the bracketing number (cf. Van der Vaart, 2000, p. 270) of a relevant functional class when we study the uniform estimation error of $\hat{p}_n(\cdot;\theta)$ relative to $p^*(\cdot;\theta)$ over $\theta \in \Theta$. 2) We approximate the solution by flexible sieves so that the approximation error disappears in first order asymptotics. 3) If $f(Y, p^*(X;\theta))$ denotes the log density of (Y_i, X_i), then the asymptotic variance of $\hat{\theta}_n$ attains the Cramér-Rao lower bound. Thus, $\hat{\theta}_n$ is asymptotically efficient.

Remark 5. In Appendix C, we examine the discrete state setting and establish consistency and asymptotic normality for both estimators; see Theorems A4 and A5 for the joint algorithm, and Theorems A7 and A8 for the nested algorithm. Compared to the continuous state setting, the discrete case does not require accounting for approximation error when using the sieve method to approximate the true solution $p^*(\cdot;\theta)$. As a result, the theoretical derivation for discrete states is considerably simpler. Notably, under mild conditions, the two estimators share the same limiting distribution, and their asymptotic variances achieve the Cramér-Rao lower bound.

Therefore, although these estimators differ from the maximum likelihood estimator of θ, they remain asymptotically efficient.

In fact, the techniques used to establish consistency and asymptotic normality for both estimators are largely the same.[11] To show consistency, the essential step is to establish the uniform law of large numbers for $M_n(\theta) = n^{-1}\sum_{i=1}^n f(Y_i, p^*(X_i;\theta))$ and a uniform bound on $|n^{-1}\sum_{i=1}^n\{f(Y_i,\hat{p}_n(X_i;\theta)) - f(Y_i,p^*(X_i;\theta))\}|$ over $\theta\in\Theta$. For asymptotic normality, we apply Theorem 5.23 of Van der Vaart (2000), as both estimators can be regarded as M-estimators. In particular, although the joint estimator $\tilde{\theta}_n$ is not defined as the maximizer of $\ell_n(p^*(\theta))$, where ℓ_n denotes the data likelihood function, we are still able to control the difference between $\ell_n(p^*(\tilde{\theta}_n))$ and $\max_{\theta\in\Theta}\ell_n(p^*(\theta))$. We show that $\tilde{\theta}_n$ nearly maximizes $\ell_n(p^*(\theta))$, which allows us to leverage the general results for M-estimators; see Section 5.2 of Van der Vaart and Wellner (1996) for details.

3.3 Discussion

MPEC: When the state space is discrete and p is finite, our method could incorporate each element of the endogenous variable p as a basis function in sieve approximation and put all of the weight on the equilibrium conditions in the inner loop. In this case, our estimator becomes the MPEC estimator.

Moreover, it is easy to show that for a given $\omega > 0$, there exists an $\epsilon > 0$ such that the optimization problem of the joint algorithm (5) has the same solution as its dual optimization problem

$$\begin{aligned}\max_{\beta,\theta}\quad & \ell(\beta,\theta)\\ s.t.\quad & \rho(\beta,\theta)\le\epsilon\end{aligned}$$

The dual problem is a natural generalization of the MPEC estimator. However, solving it numerically is challenging.

Our above-mentioned derivation also suggests a natural way to calculate standard errors for MPEC estimators.

[11] Using the same techniques, we can show that the joint estimator $\tilde{\theta}_n$, defined as in (A15) for continuous states, has the same limiting distribution under the conditions of Theorem 3.

Corollary 3.1. *When $\omega = \infty$ and $\beta = p$, the observed Fisher information can be characterized as*

$$\widehat{\mathbf{H}} = \mathbf{H}_{\theta\theta} + [\nabla\widehat{\beta}(\widehat{\theta})]'\mathbf{H}_{\beta\beta}[\nabla\widehat{\beta}(\widehat{\theta})] + [\nabla\widehat{\beta}(\widehat{\theta})]'\mathbf{H}_{\beta\theta} + [\mathbf{H}_{\beta\theta}]'\nabla\widehat{\beta}(\widehat{\theta}),$$

where $\nabla\beta(\widehat{\theta}) = -[\nabla_\beta g]^{-1} \times \nabla_\theta g$ on the right-hand side, and the matrices in bold are the four blocks in the Hessian matrix generated from a constrained maximization algorithm,

$$\begin{bmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\theta} \\ \mathbf{H}_{\theta\beta} & \mathbf{H}_{\theta\theta} \end{bmatrix}.$$

To the best of our knowledge, this result is new in the literature. Su and Judd (2012) suggest obtaining standard errors through bootstrapping. We derive the general result in Appendix D. Here, we consider $p, \beta, \theta \in \mathbb{R}$, as in the simple example with $x_j = 1$, to explain the idea. When $\omega = \infty$ and $\beta = p$, our estimator is effectively an MPEC estimator,

$$\max_{g(\beta,\theta)=0} \quad \ell(\beta, \theta).$$

The MPEC approach forms the Lagrangian function $h(\beta, \theta, \omega) = \ell(\beta, \theta) + \lambda g(\beta, \theta)$. Note that this multiplier λ should not to be confused with the smoothing parameter ω for general PSE. By definition, we have $g(\widehat{\beta}(\theta), \theta) = 0$. Its first-order and second-order derivatives are

$$\begin{aligned} g_\beta\widehat{\beta}'(\theta) + g_\theta &= 0 \\ g_{\beta\beta}[\widehat{\beta}'(\theta)]^2 + 2g_{\beta\theta}\widehat{\beta}'(\theta) + g_\beta\widehat{\beta}''(\theta) + g_{\theta\theta} &= 0, \end{aligned}$$

which allow for expressing $\widehat{\beta}'(\theta)$ and $\widehat{\beta}''(\theta)$ in the gradient of g.

On the other hand, the second-order derivative of the likelihood is

$$\begin{aligned} \widehat{\ell}_{\theta\theta}(\theta)|_{\theta=\widehat{\theta}} &= \ell_{\beta\beta}[\widehat{\beta}'(\theta)]^2 + 2\ell_{\beta\theta}\widehat{\beta}'(\theta) + \ell_\beta\widehat{\beta}''(\theta) + \ell_{\theta\theta} \\ &= \Big[\left(\frac{g_\theta}{g_\beta}\right)^2 (\ell_{\beta\beta} + \lambda g_{\beta\beta}) - 2\frac{g_\theta}{g_\beta}(\ell_{\beta\theta} + \lambda g_{\beta\theta}) + (\ell_{\theta\theta} + \lambda g_{\theta\theta}) \Big] \Big|_{\beta=\widehat{\beta}(\widehat{\theta}),\theta=\widehat{\theta}}, \end{aligned}$$

where λ denotes the associated Lagrange multiplier reported by a constrained maximization algorithm. The last equation follows from the Lagrange multiplier theorem

that $\ell_\beta + \lambda g_\beta = 0$ at the optimum $(\beta = \widehat{\beta}(\widehat{\theta}), \theta = \widehat{\theta})$ and the first-order and second-order derivatives of the equilibrium constraints. All terms on the RHS are readily available if MPEC converges. We recommend supplying the analytic gradient and Hessian, as the numerical one can be inaccurate.

To the best of our knowledge, the theoretical properties of the MPEC estimator for structural models with continuous states remain unexplored. In contrast, our proposed method offers a rigorous framework for conducting statistical inference for θ with either discrete, continuous, or both types of states. We believe this represents a critical advancement for practical applications.

Approximate MLE: We now discuss the extreme case when we let $\omega = \infty$. That is, for each guess of the model parameter θ, we find the best approximation to minimize any deviation from the equilibrium condition and then evaluate the likelihood by plugging in this best approximation. Specifically, our estimator becomes equivalent to

$$\begin{aligned} \max_{\theta} \quad & \ell(p^{\beta(\theta)}, \theta; \text{data}) \\ \text{where} \quad & \beta(\theta) = \arg\min_{\beta} \rho(p^\beta, \Psi(p^\beta, \theta)), \end{aligned}$$

which looks similar to MLE, with an important difference that we only search for the best approximation in the inner loop. We call this special case of our estimator the approximate MLE (AMLE). Such approximate solution approaches have appeared in the literature. See, e.g., Keane and Wolpin (1994, 1997) use sieves to approximate solutions in dynamic structural models.

One may wonder about the advantages of gradually changing ω instead of directly considering the limiting case. AMLE ignores the data when finding the best approximation of the solution for each θ. Because the data are informative about the true strategies p, our general sieve-based efficient estimator may perform better than AMLE. By gradually updating the smoothing parameter, we shift the weight from the data to the equilibrium condition. At the minimum, a preliminary nonparametric estimate of $\widehat{p}$ (by letting $\omega = 0$) constitutes a good starting value for the inner loop but is subject to issues with nonparametric estimates. When the smoothing parameter increases, more weight is given to the equilibrium condition. By forcing model

restrictions more strongly, the estimates converge to MLE estimates.

4 Application: Walmart versus Kmart Entry Game

In this section, we apply our methodology to an entry game between Walmart and Kmart, using a dataset published by Jia (2008). A detailed description of the industry and data is available in the original paper.

4.1 Data

The original dataset includes 2,065 markets, each representing a county with an average population ranging from 5,000 to 64,000, covering the years 1988 to 1997. For our analysis, we focus on the year 1997. The market-level variables include the log of county population (pop), the log of retail sales per capita (spc), and the percentage of urban population (urban). Walmart-specific variables include an intercept, the log of distance to Bentonville (dbenton), and an indicator for the southern region. Kmart-specific variables include an intercept and an indicator for the Midwest region (midwest). These variables capture key variations in the data. For instance, a simple scatter plot of the total number of firms shows that neither firm enters the market when SPC is too low.

Denote the data as $\{d_{\mathcal{W}m}, d_{\mathcal{K}m}, x_{\mathcal{W}m}, x_{\mathcal{K}m}, z_m\}_{m=1}^n$, where $\mathcal{W}$ and $\mathcal{K}$ represent Walmart and Kmart, respectively. Here, d_{jm} is firm j's entry decision in market m, x_{jm} includes firm-specific covariates, including a constant, and z_m contains market-specific covariates. Table 4.1 provides summary statistics for the sample used in our analysis.

Table 2: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
$d_{\mathcal{W}m}$	0.48	0.50	0	1
$d_{\mathcal{K}m}$	0.19	0.39	0	1
pop	2.98	0.67	1.54	4.37
spc	8.20	0.47	5.08	10.66
urban	0.33	0.24	0	1
dbenton	6.24	0.63	3.01	8.29
southern	0.50	0.50	0	1
midwest	0.42	0.49	0	1

4.2 Empirical Model

For the purpose of illustrating our method, we model the entry game between Walmart and Kmart as a static game with incomplete information. Two players, Walmart ($\mathcal{W}$) and Kmart ($\mathcal{K}$), decide whether to enter a market. We assume that they make independent decisions across markets. Let $d_j = 1$ if firm j is active and 0 otherwise. The payoff function of firm j depends on its own productivity, whether its competitor enters or not, market- and firm-specific covariates, and private information:

$$u_j(d_j, d_{-j}) = \underbrace{X_j'\alpha + Z'\gamma}_{\xi_j} - \Delta d_{-j} + \epsilon_{j1},$$

if $d_j = 1$ and $= \epsilon_{j0}$ otherwise, where $X = (X_{\mathcal{W}}, X_{\mathcal{K}})'$ is firm characteristics that affect only the focal firm's profit and Z is market characteristics common to both firms. For convenience, we denote $\xi_j = X_j'\alpha + Z'\gamma$.

Firm j's profit is ξ_j under monopoly and $\xi_j - \Delta$ under duopoly. Note we allow asymmetry in monopoly profit by including a constant in firm-specific covariates. The term $Z'\gamma$ is common among all firms. Denote $\theta = (\alpha, \gamma, \Delta)'$, market- and firm-specific characteristics (x, z) are common knowledge, and firm j's private information ϵ_j is type-1 extreme value distributed and independent of ϵ_{-j}.

Therefore, the probability that firm j chooses to enter is

$$p_j = \frac{1}{1 + \exp\{-\xi_j + p_{-j}\Delta\}},$$

where p_{-j} is its competitor's entry probability. Denote the CCPs as $p = (p_{\mathcal{W}}, p_{\mathcal{K}})'$. Define the best response mapping from CCP to CCP $\Psi : p \to p$. In equilibrium, we must have

$$p = \Psi(p, \theta).$$

We define the likelihood function as

$$\ell(p^\beta, \theta) = \sum_{j=\mathcal{W},\mathcal{K}} \sum_{m=1}^{n} \left\{ d_{jm} \log \left[p_j^\beta(\chi_m) \right] + (1 - d_{jm}) \log \left[1 - p_j^\beta(\chi_m) \right] \right\},$$

where $\chi_m = (x_{\mathcal{W}m}, x_{\mathcal{K}m}, z_m)'$. The approximation structure p_j^β is introduced below.

We define the approximated penalization term as

$$\widehat{\rho}(\beta,\theta) = \sum_{j=\mathcal{W},\mathcal{K}} \sum_{m=1}^{n} \left[p_j^{\beta}(\chi_m) - \Psi_j \left(p_{-j}^{\beta}(\chi_m), \theta \right) \right]^2,$$

which accounts for the equilibrium conditions for the set of observed market-specific covariates. In addition, we supply analytic gradient; see Appendix F.2.

Approximation Structure: We now consider the approximation of the CCPs. A naive approach is to approximate them as a flexible function of all market- and firm-specific covariates $p_j(x_j, x_{-j}, z)$, which is of six dimensions in our empirical setting. To ensure that the approximation error disappears in first order asymptotics, the dimension of the approximation parameter K needs to be large, leading to substantial computational challenges.

We propose a novel approximation structure that leverages the model structure: the deterministic component in the payoff function is linear in the parameters. As a result, how covariates (x, z) determine the endogenous variable p becomes a two-index model $p^*(\xi_j, \xi_{-j})$, which is much easier to approximate than a six-dimensional function. Using cubic basis functions following Luo et al. (2018), we propose to approximate the CCPs in our empirical model by

$$p^{\beta}(\xi_j, \xi_{-j}) = \sigma\left(\sum_{\imath=1}^{K} \sum_{\jmath=1}^{K} \beta_{\imath\jmath} s_{\imath}\big(\sigma(\xi_j)\big) s_{\jmath}\big(\sigma(\xi_{-j})\big) \right), \tag{11}$$

where $s_{\imath}(\cdot)$ and $s_{\jmath}(\cdot)$ are cubic spline basis functions on $[0,1]$, $\beta = (\beta_{11}, \ldots, \beta_{KK})'$ and $\sigma(\cdot) = (1 + e^{-\cdot})^{-1}$ representing the logistic function. In principle, we can use a different number of basis functions in the two dimensions. For convenience, we will use the same number K and refer to it as the approximation parameter.

Note that the logistic function appears three times but for different reasons. First, because $s_{\imath}(\cdot)$ and $s_{\jmath}(\cdot)$ are cubic spline basis functions on $[0,1]$, the inner ones $\sigma(\xi_j)$ and $\sigma(\xi_{-j})$ transforms unbounded payoff indices ξ_j and ξ_{-j} into bounded ones on $[0,1]$. Interestingly, these correspond to stand-alone entry probabilities when firms ignore strategic interaction. Second, the outer one transforms an approximation of the ex-ante value of entry $\sum_{\imath=1}^{K} \sum_{\jmath=1}^{K} \beta_{\imath\jmath} s_{\imath}(\,\cdot\,) s_{\jmath}(\,\cdot\,)$, before observing T1EV errors,

into CCPs. Altogether, our approximation structure is a hybrid of a simple neural network and a tensor product linear sieve space. It leverages the index structure in the payoff function and hence reduces the dimension of the approximation parameters needed.

Remark 6. To our knowledge, no nested fixed-point algorithm or other numerical algorithms exist for finding all equilibria in such games, rendering MLE challenging to apply. In addition, the MPEC estimator would solve a constrained maximization problem with thousands of unknowns, making it computationally difficult. Finally, NPL has no guarantee of convergence in empirical games.

4.3 Estimation Results

The algorithms proposed in Section 3.1 share the same asymptotic properties and perform similarly in simple settings. For practical, real-world applications, we recommend breaking the search process into more manageable steps. Specifically, we suggest using the joint algorithm with a small smoothing parameter to identify good starting values, followed by the alternating iterative version of the nested algorithm for the main estimation. The following algorithm summarizes the procedure.

Two-step methods are generally less efficient than MLE and also rely on consistent first-stage estimates of the CCPs. Ideally, this first-stage estimation should be nonparametric, as the functional form of the solution is unknown, even when the profit and best response functions are known. However, this leads to the well-known curse of dimensionality. To address this, we propose a novel two-step approach that leverages the single-index structure, thereby avoiding the curse of dimensionality.[12] Specifically, we first obtain the sieve MLE of the CCPs, $\widehat{p}_j$, using the same approximation structure as in Equation (11), and then estimate the parameters by maximizing the pseudo-likelihood function

$$\max_{\theta} \quad \sum_{j=\mathcal{W},\mathcal{K}} \sum_{m=1}^{n} \left[d_{jm} \log \left\{ p_j^{\theta}(\chi_m) \right\} + (1 - d_{jm}) \log \left\{ 1 - p_j^{\theta}(\chi_m) \right\} \right],$$

where $p_j^{\theta} = \frac{1}{1+\exp\{-\xi_j(\alpha,\gamma)+\widehat{p}_{-j}\Delta\}}$. Here, $\widehat{p}_{-j} = p^{\widetilde{\beta}}(\xi_{-j}(\widetilde{\alpha},\widetilde{\gamma})), \xi_j(\widetilde{\alpha},\widetilde{\gamma}))$, where $(\widetilde{\alpha},\widetilde{\gamma},\widetilde{\beta})$

[12] This approach can also be applied to the simple model in Section 2 when extending it to multiple product characteristics. Related strategies appear in the econometrics literature on single-index regression models, such as Stoker (1986) and Powell et al. (1989).

Algorithm 1 Alternating Iterative Estimation Algorithm

Input: Data $\{d_{\mathcal{W}m}, d_{\mathcal{K}m}, x_{\mathcal{W}m}, x_{\mathcal{K}m}, z_m\}_{m=1}^n$; initial value $\omega^{(1)}$
for $t = 1$ to T_{outer} **do**
    Given $\omega^{(t)}$, initialize $\theta^{(0)}$
    for $s = 1$ to T_{inner} **do**
        Given $(\omega^{(t)}, \theta^{(s-1)})$, estimate $\beta^{(s)}$ by maximizing the penalized likelihood (6)
        Given $(\omega^{(t)}, \beta^{(s)})$, estimate $\theta^{(s)}$ by maximizing the data likelihood (7)
        if $|\theta^{(s)} - \theta^{(s-1)}| <$ tolerance **then**
            Set $\theta(\omega^{(t)}) = \theta^{(s)}$; break
        end if
    end for
    if $|\theta(\omega^{(t)}) - \theta(\omega^{(t-1)})| <$ tolerance **then**
        break
    else
        Update $\omega^{(t+1)} = L \times \omega^{(t)}$
    end if
end for
Output: Final estimates $\widehat{\theta} = \theta(\omega^{(t)})$

denotes the first-step estimates of the parameters (α, γ, β).

Table 4.3 shows the estimated parameters when the number of basis functions, K, varies from 10 to 30. The second last column reports the maximum likelihood estimates assuming equilibrium uniqueness.[13] The last column reports the two-step estimates.[14] The estimates are quite similar across different estimators. All estimates are significant at the 5% level and their signs are consistent with Jia (2008). More populated areas tend to have more stores, and higher retail sales per capita predict increased entry. Urbanized areas also attract more entry. The southern region dummy variable and the log of the distance to Walmart's headquarters in Bentonville, Arkansas, both significantly predict Walmart's entry decisions. Similarly, because Kmart's headquarters are located in Troy, Michigan, the dummy variable for the Midwest region is predictive of Kmart's entry decisions.

The maximum likelihood estimates imply that the mean and standard deviation

[13] We conduct fine grid search to check equilibrium uniqueness and find that the equilibrium is unique in each market.

[14] All simulations and the empirical application in this paper were implemented in Matlab R2024b, using `fminunc` with its built-in quasi-Newton algorithm wherever maximization or minimization was required, and were run on a machine with an 11th Gen Intel Core i7-11800H 2.30 GHz processor and 16 GB of RAM.

Table 3: Estimation Results

	SEES			MLE	2-Step
K	10	20	30		
Market-specific					
pop	3.38	3.37	3.34	3.29 (0.15)	3.40
spc	2.81	2.83	2.82	2.80 (0.19)	2.99
urban	2.37	2.39	2.37	2.32 (0.31)	2.50
Walmart-specific					
intercept	-22.90	-23.03	-22.77	-22.29 (1.73)	-24.33
dbenton	-1.86	-1.85	-1.87	-1.90 (0.13)	-1.87
south	1.02	1.02	1.06	1.10 (0.15)	1.07
Kmart-specific					
intercept	-36.21	-36.31	-36.22	-35.99 (1.69)	-37.70
midwest	0.66	0.66	0.65	0.65 (0.14)	0.58
Δ	1.96	1.98	1.85	1.65 (0.27)	2.12

Note: The model is estimated using the proposed method with $K = 10, 20, 30$, MLE, and the two-step method. The SEES approach results in final ω values of 10^3, 10^4, and 10^5, with corresponding ρ values of 0.1679, 0.0247, and 0.0180, respectively.

of $\xi_{\mathcal{W}}$ are -0.11 and 3.43, respectively, while the mean and standard deviation of $\xi_{\mathcal{K}}$ are -2.20 and 3.30, respectively. The large coefficients on firm dummies suggest substantial entry costs. Note that Walmart is a dominant firm with a penetration rate of 48%, while Kmart is relatively weak with a penetration rate of 19%. This explains the much lower coefficient on the Kmart firm dummy. The proposed sieve estimator performs well across different K. The larger the approximation parameter K is, the closer the estimates become to the maximum likelihood estimates. The estimates from the proposed two-step estimator have larger biases.

5 Conclusion

A structural model is based on economic theory and describes how endogenous variables relate to a set of explanatory variables. This relationship is often expressed as an implicit function dependent on unknown parameters, which can be costly to solve. Two-step methods avoid solving the model but rely heavily on the accuracy of the first-step nonparametric estimation. We introduce SEES as a new class of

estimators that use a sieve to approximate the solution while penalizing deviations from the equilibrium condition. SEES are straightforward to apply, at least as fast as alternative approaches like MLE, and more robust across various models. We believe our method will become a valuable tool in structural estimation.

Appendix

A Optimal Monopoly Pricing With Logit Demand

In this section, we derive Equation (2). Rearranging terms gives $\xi_j - \alpha P_j + \exp(\xi_j - \alpha P_j) = \xi_j - \alpha c_j - 1$, which reduces to

$$\xi_j(x) - P_j^*(x) + \exp\{\xi_j(x) - P_j^*(x)\} = \xi_j(x) - 1$$

under our assumptions $c_j = 0, \alpha = 1$. Denote $M(x) = \exp\{\xi_j(x) - P_j^*(x)\}$. The FOC can be rewritten as $\log M(x) + M(x) = \xi_j(x) - 1$. Therefore, we have

$$M(x) = W(\exp\{\xi_j(x) - 1\}),$$

applying an alternative definition of the Lambert W function $\log W(v) + W(v) = \log v$. That is, the optimal price satisfies

$$P_j^*(x) = \xi_j(x) - \log M(x) = \xi_j(x) - \{\xi_j(x) - 1 - M(x)\} = 1 + W(\theta x),$$

where the first equation follows the definition of $M(\cdot)$, the second equation follows the rewritten FOC, and the last equation follows $\xi_j(x) = \log x + \log \theta + 1$.

B Proofs in Section 3.2

In this section, we focus on developing consistency and asymptotic normality for the estimators of θ obtained from the nested algorithm in the setting of continuous states. To this end, we first establish an approximation result that helps us to determine an appropriate choice of the sieve space $\mathcal{B}_n$.

Lemma A1. *Under Assumptions 2 and 3, we have when* $|\tau^{(n)}| = o(1)$,

$$\sup_{\theta \in \Theta} \inf_{\eta \in \mathcal{B}_n} [\|p^*(\cdot;\theta) - \eta(\cdot)\|_\infty \vee \|\Psi(p^*(\cdot;\theta);\theta) - \Psi(\eta(\cdot),\theta)\|_\infty] = O(|\tau^{(n)}|^4),$$

where $a \vee b = \max(a, b)$ *denotes the maximum of two real numbers* a *and* b.

Proof of Lemma A1. Under Assumption P4, Ψ has continuous fourth order derivatives. Therefore, $\mathrm{d}^4 p^*(x;\theta)/\mathrm{d}x^4$ is a continuous function of x and θ for $(x, \theta) \in$

$[0,T]\times\Theta$. Because Θ is compact, it follows that

$$\sup_{\theta\in\Theta}\left\|\frac{\mathrm{d}^4 p^*(x;\theta)}{\mathrm{d}x^4}\right\|_\infty<\infty.$$

Letting $|\tau^{(n)}|\to 0$ as $n\to\infty$, by Theorem 2 and Theorem 4 in Hall and Meyer (1976), we have

$$\begin{aligned}
&\sup_{\theta\in\Theta}\inf_{\eta_\theta\in\mathcal{B}_n}\|p(\cdot;\theta)-\eta_\theta\|_\infty\le C_0\sup_{\theta\in\Theta}\left\|\frac{\mathrm{d}^4 p(x;\theta)}{\mathrm{d}x^4}\right\|_\infty|\tau^{(n)}|^4\to 0,\\
&\sup_{\theta\in\Theta}\inf_{\eta_\theta\in\mathcal{B}_n}\|\Psi(p^*(\cdot;\theta);\theta)-\Psi(\eta_\theta(\cdot),\theta)\|_\infty\\
&\qquad\le\sup_{|p|\le r,\theta\in\Theta}\left|\frac{\mathrm{d}\Psi}{\mathrm{d}p}(p;\theta)\right|\times\sup_{\theta\in\Theta}\inf_{\eta_\theta\in\mathcal{B}_n}\|p^*(\cdot;\theta)-\eta_\theta\|_\infty,
\end{aligned}\tag{A1}$$

for some positive constant C_0 and r. On the right-hand side of (A1), since we choose η_θ that best approximates $p^*(x;\theta)$ over $\mathcal{B}_n$ and Θ is compact, we can choose a sufficient large but finite r such that we restrict our attention to p which is bounded by r from above. Thus, the left-hand side of the bottom line of (A1) is of order $|\tau^{(n)}|^4$ as well. The proof is completed. □

Proof of Theorem 1. Let $\rho(p,\Psi(p,\theta))=\int_0^T[p(x)-\Psi(p(x),\theta)]^2\,\mathrm{d}x$. Based on the definition of $\hat{p}_n(\cdot,\theta)$, we have for any $\theta\in\Theta$,

$$\ell_n(\hat{p}_n(\cdot;\theta))-\omega\rho(\hat{p}_n(\cdot;\theta),\Psi(\hat{p}_n(\cdot;\theta),\theta))\ge\ell_n(p_{\theta,n}(\cdot))-\omega\rho(\hat{p}_n(\cdot;\theta),\Psi(\hat{p}_n(\cdot;\theta),\theta)).$$

As $\ell_n(\hat{p}_n(\cdot;\theta))\le 0$ by Assumption 5, it follows that

$$-\omega\rho(\hat{p}_n(\cdot;\theta),\Psi(\hat{p}_n(\cdot;\theta),\theta))\ge\ell_n(p_{\theta,n}(\cdot))-\omega\rho(\hat{p}_n(\cdot;\theta),\Psi(\hat{p}_n(\cdot;\theta),\theta)).$$

Then

$$
\begin{aligned}
&\rho(\hat{p}_n(\cdot;\theta), \Psi(\hat{p}_n(\cdot;\theta),\theta)) \\
&\le -\omega^{-1}\ell_n(p_{\theta,n}(\cdot)) + \rho(\hat{p}_n(\cdot;\theta), \Psi(\hat{p}_n(\cdot;\theta),\theta)) \\
&= \omega^{-1}\mathbb{P}_n[-f(Y, p_{\theta,n}(X))] + \|p_{\theta,n}(\cdot) - \Psi(p_{\theta,n}(\cdot),\theta)\|^2_{L^2[0,T]} \\
&= \omega^{-1}\mathbb{P}_n[-f(Y, p_{\theta,n}(X))] + \|p_{\theta,n}(\cdot) - p^*(\cdot;\theta) + \Psi(p^*(\cdot;\theta),\theta) - \Psi(p_{\theta,n}(\cdot),\theta)\|^2_{L^2[0,T]} \\
&\le \omega^{-1}\mathbb{P}_n[-f(Y, p_{\theta,n}(X))] + 2\|p_{\theta,n}(\cdot) - p^*(\cdot;\theta)\|^2_{L^2[0,T]} \\
&\qquad + 2\|\Psi(p^*(\cdot;\theta),\theta) - \Psi(p_{\theta,n}(\cdot),\theta)\|^2_{L^2[0,T]} \\
&\le \omega^{-1}\mathbb{P}_n[-f(Y, p_{\theta,n}(X))] + 2\|p_{\theta,n}(\cdot) - p^*(\cdot;\theta)\|^2_{\infty} \\
&\qquad + 2\|\Psi(p^*(\cdot,\theta),\theta) - \Psi(p_{\theta,n}(\cdot),\theta)\|^2_{\infty} \\
&\le \omega^{-1}\mathbb{P}_n[-f(Y, p_{\theta,n}(X))] + C_0 r_n^2.
\end{aligned}
$$

for some constant C_0, where the last inequality holds by Lemma A1.

Note that $\sup_{\theta\in\Theta}\|p_{\theta,n}(\cdot)\|_\infty \le r$ by the definition of the sieve space $\mathcal{B}_n(r)$. By invoking the same argument for proving equation (A14), we have

$$\sup_{\theta\in\Theta}\mathbb{P}_n[-f(Y, p_{\theta,n}(X))] = O_{\mathbb{P}}(1).$$

It follows that

$$
\begin{aligned}
\sup_{\theta\in\Theta}\rho(\hat{p}_n(\cdot;\theta), \Psi(\hat{p}_n(\cdot;\theta),\theta)) &\le \omega^{-1}O_{\mathbb{P}}(1) + C_0 r_n^2 \\
&= O_{\mathbb{P}}(\omega^{-1}) + O(r_n^2).
\end{aligned}
$$

As $\rho(p, \Psi(p,\theta)) = \int_0^T [p(x) - \Psi(p(x),\theta)]^2\,\mathrm{d}x$, we have

$$\sup_{\theta\in\Theta}\|\hat{p}_n(\cdot;\theta) - \Psi(\hat{p}_n(\cdot;\theta),\theta)\|^2_{L^2[0,T]} \le O_{\mathbb{P}}(\omega^{-1}) + O(r_n^2).$$

Since $p^*(x;\theta)$ satisfies $p(x) = \Psi(p(x),\theta)$,

$$\hat{p}_n(x;\theta) - \Psi(\hat{p}_n(x;\theta),\theta) = g(p^*(x;\theta),\theta) - g(\hat{p}_n(x;\theta),\theta)$$

holds for for any $x \in [0,T]$ and $\theta \in \Theta$. Moreover, under Assumption 3, we have

$$\begin{aligned}
&\sup_{\theta\in\Theta} \quad \|\hat{p}_n(\cdot;\theta) - p^*(\cdot;\theta)\|^2_{L^2[0,T]} \\
&\leq \sup_{\theta\in\Theta} C_g^2 \|g(p^*(\cdot;\theta),\theta) - g(\hat{p}_n(\cdot;\theta),\theta)\|^2_{L^2[0,T]} \\
&= \sup_{\theta\in\Theta} C_g^2 \|\hat{p}_n(\cdot;\theta) - \Psi(\hat{p}_n(\cdot;\theta),\theta)\|^2_{L^2[0,T]} \\
&= O_{\mathbb{P}}(\omega^{-1}) + O(r_n^2).
\end{aligned}$$

Lastly, under Assumption 1, one obtains

$$\sup_{\theta\in\Theta} \|\hat{p}_n(\cdot;\theta) - p^*(\cdot;\theta)\|^2_{L^2(\mathbb{P})} = O_{\mathbb{P}}(\omega^{-1}) + O(r_n^2).$$

The proof is completed.

□

To apply Theorem 5.23 of Van der Vaart (2000) to show asymptotic normality for $\hat{\theta}_n$, we need the following lemma.

Lemma A2. *Under the same conditions in Theorem 3, we have*

$$\ell_n(p^*(\cdot;\hat{\theta}_n)) \geq \sup_{\theta\in\Theta} \ell_n(p^*(\cdot;\theta)) - o_{\mathbb{P}}(n^{-1}), \tag{A2}$$

and

$$\sup_{\theta\in\Theta} |\ell_n(p^*(\cdot;\theta)) - M(\theta)| = o_{\mathbb{P}}(1). \tag{A3}$$

Proof of Lemma A2. Under Assumption 6, there exists a constant $\delta > 0$, such that

$$\sup_{|p|\leq r+1} \left|\frac{\partial f}{\partial p}(y,p)\right| \leq \delta^{-1} \left\{1 + \inf_{|p|\leq r+1} \left|\frac{\partial f}{\partial p}(y,p)\right|\right\}. \tag{A4}$$

Then it follows

$$
\begin{aligned}
&|\ell_n(\hat{p}_n(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} |f(Y_i, \hat{p}_n(X_i;\theta)) - f(Y_i, p^*(X_i;\theta))| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\{ \sup_{|p|\leq r+1} \left| \frac{\partial f}{\partial p}(Y_i, p) \right| \right\} |\hat{p}_n(X_i;\theta) - p^*(X_i;\theta)| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\delta} \left\{ 1 + \inf_{|p|\leq r+1} \left| \frac{\partial f}{\partial p}(Y_i, p) \right| \right\} |\hat{p}_n(X_i;\theta) - p^*(X_i;\theta)| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\{ 1 + \left| \frac{\partial f}{\partial p}(Y_i, p^*(X_i;\theta_0)) \right| \right\} |\hat{p}_n(X_i;\theta) - p^*(X_i;\theta)|.
\end{aligned}
\tag{A5}
$$

Next we identify the stochastic order of the right-hand side of (A5) using empirical process theory. To simplify the notation, for any function $\eta \in \mathcal{B}_n(r)$ and $\theta \in \Theta$, we denote $H(r, \eta, \theta)$ as the right-hand side function from one single observation with parameter (η, θ), i.e.,

$$
H(r, \eta, \theta) = \left\{ 1 + \left| \frac{\partial f}{\partial p}(y, p^*(x;\theta_0)) \right| \right\} |\eta(x) - p^*(x;\theta)|.
$$

Then we define

$$
\mathbb{G}_n[H(R, \hat{p}_n(\cdot;\theta), \theta)] = \sqrt{n}(\mathbb{P}_n - \mathbb{P})[H(R, \hat{p}_n(\cdot;\theta), \theta)].
$$

To find the upper bound on $\mathbb{G}_n[H(R, \hat{p}_n(\cdot;\theta), \theta)]$, we consider a function class $\mathcal{L}_n$ defined by $\{H(r, \eta, \theta) : \eta \in \mathcal{B}_n(r), \theta \in \Theta\}$. By Assumption 6 and the definition of the sieve space $\mathcal{B}_n(r)$, the class $\mathcal{L}_n$ has an upper bound $O_{\mathbb{P}}(\log n)$. Additionally, this class can be treated as a class of functions indexed by θ and $\{\beta_j\}_{j=1}^{K}$, which are the B-spline coefficients of η in $\mathcal{B}_n(r)$. It is straightforward to verify that functions in $\mathcal{L}_n$ are Lipschitz continuous with respect to all parameters and the Lipschitz constant is bounded by $O_{\mathbb{P}}(\log n)$. Additionally, since θ is bounded by some constant, and all these B-spline coefficients are bounded by r, they must reside in a hypercube of $\mathbb{R}^{K+1}$. Hence, if we partition this large hypercube into a set of smaller hypercubes with scale length ϵ, the cardinality number of this set is no more than $O(\epsilon^{-K})$. Furthermore, by the Lipschitz property of functions in $\mathcal{L}_n$, the L_∞ distance between any two functions in the same

subcube is bounded by $O_{\mathbb{P}}(\log n)K\epsilon$. Therefore, the bracketing number (cf. Van der Vaart, 2000, p. 270) of $\mathcal{L}_n$ satisfies $N_{[\cdot]}(O_{\mathbb{P}}(\log n)K\epsilon, \mathcal{L}_n, L_\infty) \leq O(1)\epsilon^{-K_n}$. Then by Theorem 19.35 of Van der Vaart (2000) and $n^{1/4} = o(K)$, we have in probability

$$\begin{aligned}\sqrt{n}\mathbb{E}^*\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_n} &\leq O_{\mathbb{P}}(1)\int_0^1 \sqrt{\log\left\{\frac{2(\log n)K}{\epsilon}\right\}^{2K}}\,\mathrm{d}\epsilon \\ &\leq O_{\mathbb{P}}(1)K^{1/2}\sqrt{\log(K)}.\end{aligned}$$

Therefore, $\mathbb{G}_n[H(R, \hat{p}_n(\cdot;\theta), \theta)]$ is bounded by $O_{\mathbb{P}}(K^{1/2}\sqrt{\log(K)}/\sqrt{n})$ from above, which is $o_{\mathbb{P}}(1)$ by the assumption $K^2\log(K) = o(n)$.

Furthermore, with some abuse of notation, we still use $\mathcal{L}_n$ to denote the class

$$\left\{H(r,\theta) : H(r,\theta) = \left\{1 + \left|\frac{\partial f}{\partial p}(y, p^*(x;\theta_0))\right|\right\}|\hat{p}_n(x;\theta) - p^*(x;\theta)|\right\}.$$

Obviously, the index set Θ is totally bounded when equipped with the Euclidean distance. Moreover, we can choose $F_n = (2r+1)\{1 + |\partial f(y, p^*(x;\theta_0)/\partial p\}$ to be the envelop function of $\mathcal{L}_n$, which has finite moments of all orders. Additionally, from the preceding arguments, we know this class is equi-continuous with respect to θ. Therefore, by Theorem 2.11.23 of Van der Vaart and Wellner (1996), $\mathbb{G}_n[H(R, \hat{p}_n(\cdot;\theta), \theta)]$ is bounded by $o_{\mathbb{P}}(1/n)$ from above. Consequently, it follows from Theorem 1 that

$$\sup_{\theta\in\Theta}|\ell_n(\hat{p}_n(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))| = o_{\mathbb{P}}(n^{-1}) \tag{A6}$$

if $r_n = o(n^{-1})$ and $\omega/n^2 \to \infty$.

By definition of $p^*(\cdot;\theta)$ and $\hat{\theta}_n$, we have

$$\begin{aligned}&\ell_n(p^*(\cdot;\hat{\theta}_n)) \\ &\geq \ell_n(\hat{p}_n(\cdot;\hat{\theta}_n)) - \sup_{\theta\in\Theta}|\ell_n(\hat{p}_n(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))| \\ &\geq \sup_{\theta\in\Theta}\ell_n(p^*(\cdot;\theta)) - 2\sup_{\theta\in\Theta}|\ell_n(\hat{p}_n(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))|.\end{aligned}$$

Hence, the relation in (A2) holds. Equation (A3) can be obtained with a slight modification of Lemma A6 in Appendix C.1. □

Next we prove Theorem 2.

Proof of Theorem 2. By (A3) and (A6), we have

$$\begin{aligned}
&M(\theta_0) - M(\hat{\theta}_n) \\
&= M_n(\theta_0) - M_n(\hat{\theta}_n) + o_{\mathbb{P}}(1) \quad \text{by (A3)} \\
&= \ell_n(p^*(\cdot;\theta_0)) - \ell_n(p^*(\cdot;\hat{\theta}_n)) + o_{\mathbb{P}}(1) \\
&\leq \ell_n(\hat{p}_n(\cdot;\theta_0)) + \sup_{\theta\in\Theta} |\ell_n(\hat{p}(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))| - \ell_n(\hat{p}(\cdot;\hat{\theta}_n)) \\
&\quad + \sup_{\theta\in\Theta} |\ell_n(\hat{p}(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))| + o_{\mathbb{P}}(1) \\
&= \ell_n(\hat{p}(\cdot;\theta_0)) - \ell_n(\hat{p}(\cdot;\hat{\theta}_n)) + 2\sup_{\theta\in\Theta} |\ell_n(\hat{p}(\cdot;\theta)) - \ell_n(p^*(\cdot;\theta))| + o_{\mathbb{P}}(1) \\
&\leq \ell_n(\hat{p}(\cdot;\theta_0)) - \ell_n(\hat{p}(\cdot;\hat{\theta}_n)) + o_{\mathbb{P}}(1). \quad \text{by (A6)}
\end{aligned}$$

By definition of $\hat{\theta}_n$, we have

$$\ell_n(\hat{p}(\cdot;\hat{\theta}_n)) \geq \ell_n(\hat{p}(\cdot;\theta_0)),$$

so $M(\theta_0) - M(\hat{\theta}_n) \leq o_{\mathbb{P}}(1)$. By (A20), we have

$$\{d(\hat{\theta}_n, \theta_0) \geq \delta\} \subset \{M(\theta_0) - M(\hat{\theta}_n) \geq \gamma\} \subset \{o_{\mathbb{P}}(1) \geq \gamma\}.$$

Therefore,

$$\mathbb{P}_{\theta_0}(d(\hat{\theta}_n, \theta_0) \geq \delta) \leq \mathbb{P}_{\theta_0}(o_{\mathbb{P}}(1) \geq \gamma),$$

which converges to 0 as n approaches infinity. As δ is an arbitrary positive number, $\hat{\theta}_n$ is a consistent estimator of θ_0. This completes the proof. □

Now we are ready to prove Theorem 3.

Proof of Theorem 3. We follow Theorem 5.23 of Van der Vaart (2000) to prove asymptotic normality of $\hat{\theta}_n$. Firstly, as we have shown in the proof of Lemma A6, under Assumption 6, we have

$$\begin{aligned}
&|f(y, p^*(x;\theta_1)) - f(y, p^*(x;\theta_2))| \\
&\leq \frac{1}{\eta}\left[1 + \left|\frac{\partial f}{\partial p}(y, p^*(x;\theta_0))\right|\right] |p^*(\theta_1) - p^*(\theta_2)| \\
&\leq \frac{C}{\eta}\left[1 + \left|\frac{\partial f}{\partial p}(y, p^*(x;\theta_0))\right|\right] \|\theta_1 - \theta_2\|_2 \qquad \text{(A7)}
\end{aligned}$$

for some constant C and η. By Assumption 6, the right-hand side of (A7) has a finite second moment.

Next, we consider a second-order Taylor expansion for

$$M(\theta) = \mathrm{E}_{\theta_0}[f(Y, p^*(X;\theta))]$$

in a neighbourhood of θ_0. Obviously,

$$\begin{aligned}
& f(y, p(x;\theta)) \\
&= f(y, p^*(x;\theta_0)) + \left[\frac{\partial f(y, p^*(x;\theta_0))}{\partial p}\right]' \left(\frac{\partial p^*(x;\theta_0)}{\partial \theta}\right)' (\theta - \theta_0) \\
&\quad + \frac{1}{2}(\theta - \theta_0)' \left[\frac{\partial f(y, p^*(x;\theta_0))}{\partial p} \frac{\partial^2 p^*(x;\theta_0)}{\partial \theta \partial \theta'}\right. \\
&\quad \left. + \frac{\partial^2 f(y, p^*(x;\theta_0))}{\partial p^2} \left(\frac{\partial p^*(x;\theta_0)}{\partial \theta}\right)\left(\frac{\partial p^*(x;\theta_0)}{\partial \theta}\right)'\right] \\
&\quad \times (\theta - \theta_0) + R,
\end{aligned} \tag{A8}$$

where R is the remainder term. Define

$$\begin{aligned}
D(y, x, \theta) = & \frac{\partial f(y, p^*(x;\theta_0))}{\partial p} \frac{\partial^2 p^*(x;\theta_0)}{\partial \theta \partial \theta'} \\
& + \frac{\partial^2 f(y, p^*(x;\theta_0))}{\partial p^2} \left(\frac{\partial p^*(x;\theta_0)}{\partial \theta}\right)\left(\frac{\partial p^*(x;\theta_0)}{\partial \theta}\right)'.
\end{aligned}$$

Then the reminder term can be rewritten as

$$R = (\theta - \theta_0)' \left[\int_0^1 [D(y, x, \theta_0 + s(\theta - \theta_0)) - D(y, x, \theta_0)](1 - s)\,\mathrm{d}s\right](\theta - \theta_0).$$

Note that $D(y, x, \theta)$ is a $d \times d$ matrix. For any (a, b)th entry in $D(y, x, \theta)$, we can show that, for any $\theta \in \Theta$,

$$\begin{aligned}
& |D_{ab}(y, x, \theta)| \\
&\leq \left|\frac{\partial f(y, p^*(x;\theta))}{\partial p} \frac{\partial^2 p^*(x;\theta)}{\partial \theta_a \partial \theta_b}\right| + \left|\frac{\partial^2 f(y, p^*(x;\theta))}{\partial p^2} \left(\frac{\partial p^*(x;\theta)}{\partial \theta_a}\right)' \left(\frac{\partial p^*(x;\theta)}{\partial \theta_b}\right)\right| \\
&\leq \frac{C'}{\eta}\left[1 + \left|\frac{\partial f}{\partial p}(y, p^*(x;\theta_0))\right| + \left|\frac{\partial^2 f}{\partial p^2}(y, p^*(x;\theta_0))\right|\right],
\end{aligned}$$

where C' is a positive constant. Additionally, under Assumption 6, the right-hand side of the above inequality has a finite mean. Therefore, applying the dominated convergence theorem, we have

$$\mathrm{E}_{\theta_0}\left[\int_0^1 [D(y,x,\theta_0+s(\theta-\theta_0))-D(y,x,\theta_0)](1-s)\,\mathrm{d}s\right]\to 0$$

as $\theta\to\theta_0$. Then, by the Taylor expansion of $f(y,p^*(x;\theta))$, it follows that

$$M(\theta)=M(\theta_0)+\frac{1}{2}(\theta-\theta_0)'V_{\theta_0}(\theta-\theta_0)+o\left(\|\theta-\theta_0\|_2^2\right). \tag{A9}$$

Recall that $f(y,p)$ is the log density of Y_i. Therefore, there is no linear form in (A9) and the expected value of $D(Y,X,\theta_0)$ is given by V_{θ_0} as the expectation of (A22) is zero.

Finally, (A2) holds by Lemma A2. Moreover, we can easily show $\hat{\theta}_n$ is a consistent estimator of θ_0 from (A3) in Lemma A2. Combining (A7), (A9) and (A2), it follows from Theorem 5.23 in Van der Vaart (2000) that $\sqrt{n}(\hat{\theta}_n-\theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1}\mathrm{E}_{\theta_0}\left[\left\{\frac{\partial f(Y,p^*(X;\theta_0))}{\partial p}\right\}^2\left(\frac{\partial p^*(X;\theta_0)}{\partial\theta}\right)'\left(\frac{\partial p^*(X;\theta_0)}{\partial\theta}\right)\right]V_{\theta_0}^{-1},$$

if V_{θ_0} is non-singular. This completes the proof. □

Lastly, we prove Corollary 3.1.

Proof of Corollary 3.1. Denote the jth function of $g(\widehat{\beta}(\theta),\theta)=0$ as g^j. Taking its first-order derivative gives

$$\sum_k g^j_{\beta_k}\frac{\partial\beta_k}{\partial\theta_\ell}+g^j_{\theta_\ell}=0,$$

which can be written in matrix form $\nabla_\beta g\times\nabla\beta(\theta)+\nabla_\theta g=0$. Therefore,

$$\nabla\beta(\theta)=-\left[\nabla_\beta g\right]^{-1}\times\nabla_\theta g.$$

Taking its second-order derivative gives

$$\sum_k \left\{ \left[\sum_{k'} g^j_{\beta_k \beta_{k'}} \frac{\partial \beta_{k'}}{\partial \theta_{\ell'}} + g^j_{\beta_k \theta_{\ell'}} \right] \frac{\partial \beta_k}{\partial \theta_\ell} + g^j_{\beta_k \theta_\ell} \frac{\partial \beta_k}{\partial \theta_{\ell'}} \right\} + \sum_k g^j_{\beta_k} \frac{\partial^2 \beta_k}{\partial \theta_\ell \partial \theta_{\ell'}} + g^j_{\theta_\ell \theta_{\ell'}} = 0. \tag{A10}$$

Now, consider the Hessian of the likelihood function $\widehat{\ell}(\theta) = \ell(\widehat{\beta}(\theta), \theta)$,

$$\begin{aligned}
\frac{\partial^2 \widehat{\ell}(\theta)}{\partial \theta_\ell \theta_{\ell'}} =& \sum_k \left\{ \left[\sum_{k'} \ell_{\beta_k \beta_{k'}} \frac{\partial \beta_{k'}}{\partial \theta_{\ell'}} + \ell_{\beta_k \theta_{\ell'}} \right] \frac{\partial \beta_k}{\partial \theta_\ell} + \ell_{\beta_k} \frac{\partial^2 \beta_k}{\partial \theta_\ell \partial \theta_{\ell'}} + \ell_{\beta_k \theta_\ell} \frac{\partial \beta_k}{\partial \theta_{\ell'}} \right\} + \ell_{\theta_\ell \theta_{\ell'}} \\
=& \ell_{\theta_\ell \theta_{\ell'}} + \sum_k \sum_{k'} \ell_{\beta_k \beta_{k'}} \frac{\partial \beta_{k'}}{\partial \theta_{\ell'}} \frac{\partial \beta_k}{\partial \theta_\ell} + \sum_k \left[\ell_{\beta_k \theta_{\ell'}} \frac{\partial \beta_k}{\partial \theta_\ell} + \ell_{\beta_k \theta_\ell} \frac{\partial \beta_k}{\partial \theta_{\ell'}} \right] \\
& - \sum_k \sum_j \lambda_j g^j_{\beta_k} \frac{\partial^2 \beta_k}{\partial \theta_\ell \partial \theta_{\ell'}} \\
=& \underbrace{\left[\ell_{\theta_\ell \theta_{\ell'}} + \sum_j \lambda_j g^j_{\theta_\ell \theta_{\ell'}} \right]}_{\mathbf{H}_{\theta\theta}} + \sum_k \sum_{k'} \underbrace{\left[\ell_{\beta_k \beta_{k'}} + \sum_j \lambda_j g^j_{\beta_k \beta_{k'}} \right]}_{\mathbf{H}_{\beta\beta}} \frac{\partial \beta_{k'}}{\partial \theta_{\ell'}} \frac{\partial \beta_k}{\partial \theta_\ell} \\
& + \sum_k \underbrace{\left[\ell_{\beta_k \theta_{\ell'}} + \sum_j \lambda_j g^j_{\beta_k \theta_{\ell'}} \right]}_{\mathbf{H}_{\beta\theta}} \frac{\partial \beta_k}{\partial \theta_\ell} + \sum_k \underbrace{\left[\ell_{\beta_k \theta_\ell} + \sum_j \lambda_j g^j_{\beta_k \theta_\ell} \right]}_{\mathbf{H}_{\theta\beta}} \frac{\partial \beta_k}{\partial \theta_{\ell'}}
\end{aligned}$$

where the first equation follows from the Lagrange multiplier theorem, i.e., $\ell_{\beta_k} + \sum_j \lambda_j g^j_{\beta_k} = 0$, and the second equation follows from (A10). In its matrix form, we can construct the observed Fisher information,

$$\widehat{\mathbf{H}} = \mathbf{H}_{\theta\theta} + [\nabla \widehat{\beta}(\widehat{\theta})]' \mathbf{H}_{\beta\beta} [\nabla \widehat{\beta}(\widehat{\theta})] + [\nabla \widehat{\beta}(\widehat{\theta})]' \mathbf{H}_{\beta\theta} + [\mathbf{H}_{\beta\theta}]' \nabla \widehat{\beta}(\widehat{\theta}),$$

where $\nabla \beta(\widehat{\theta}) = -\left[\nabla_\beta g\right]^{-1} \times \nabla_\theta g$, and the matrices in bold are the four blocks in the Hessian matrix generated from a constrained maximization algorithm,

$$\begin{bmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\theta} \\ \mathbf{H}_{\theta\beta} & \mathbf{H}_{\theta\theta} \end{bmatrix}.$$

□

C Asymptotic Properties in Discrete State Settings

As mentioned in the main text, our proposed method can handle both continuous states, which result in an infinite-dimensional endogenous variable p, and discrete states, which lead to a finite-dimensional p. In the setting of discrete states, the approximation can be perfect, i.e., $\beta = p$. That is, $s_k(x) = \mathbb{1}(x = p_k)$, where $\mathbb{1}(\cdot)$ is the indicator function and p_k is the kth element in the endogenous variable p.

Let $p = (p_1, \ldots, p_J)' \in \mathbb{R}^J$ be the endogenous variable in (1). Without loss of generality, we assume that $\theta \in \mathbb{R}^d$ and Θ denotes the space of θ. Under certain conditions that are specified below, we establish consistency and asymptotic normality of the joint estimator of θ first. Then we develop consistency and asymptotical normality for the nested estimator of θ. In fact, the two estimators have the same asymptotic distribution.

Assumption P1. The parameter space Θ is a compact subset of $\mathbb{R}^d$.

For any given $\theta \in \Theta$, we aim to maximize the following function with respect to p

$$\ell_n(p) - \omega \|p - \Psi(p, \theta)\|_2^2, \tag{A11}$$

where ℓ_n denotes the log-likelihood corresponding to n i.i.d. observations and $\|\cdot\|_2$ is the Euclidean norm of a vector. Suppose $Y_1, \ldots, Y_n$ are i.i.d. observations, taking values in $\mathbb{R}^{d_Y}$. We assume that the likelihood function in (A11) can be written as

$$\ell_n(p) = \frac{1}{n} \sum_{i=1}^{n} f(Y_i, p),$$

where f is a function defined on $\mathbb{R}^{d_Y} \times \mathbb{R}^J$. For simplicity, we assume that $d_Y = 1$. In this context, $f(y, p)$ is actually the log density function of Y_i.

Assumption P2. There exists a compact and convex set $\Lambda \subset \mathbb{R}^J$ such that p must lie in Λ.

By Brouwer's fixed-point theorem, there must exist a solution to (1) for any $\theta \in \Theta$. For instance, $p \in [0, 1]$ represents CCP in dynamic games. Define $g(p, \theta) = \Psi(p, \theta) - p$. Obviously, the solution to Equation (1), denoted as $p^*(\theta)$, satisfies $g(p^*(\theta), \theta) = 0$. We impose the following regularity condition on $g(p, \theta)$.

Assumption P3. There exists a positive constant r such that for any $(p_1, \theta), (p_2, \theta) \in \Lambda \times \Theta$ satisfying $\|g(p_1, \theta)\|_2 \vee \|g(p_2, \theta)\|_2 \leq r$, where $a \vee b$ denotes the larger value of a and b, we have $\|p_1 - p_2\|_2 \leq C\|g(p_1, \theta) - g(p_2, \theta)\|_2$ for some constant $C > 0$.

Assumption P3 can be understood as a local inverse Lipschitz condition. Consider the Lambert function $p = W(\theta)$, which is defined implicitly by $pe^p = \theta$. In correspondence, $g(p, \theta) = \theta e^{-p} - p$. Thus, $(p+g)e^{p+g} = \theta e^g$, which implies that $p + g = W(\theta e^g)$ or

$$p(g) = -g + W\left(\theta e^g\right).$$

Since W is a continuously differentiable function by the implicit function theorem, for any g_1, g_2 satisfying $|g_1| \vee |g_2| \leq r$ with some constant r, we have $|p(g_1) - p(g_2)| \leq C|g_1 - g_2|$ for some constant C.

Assumption P4. Ψ is twice differentiable in both p and θ, and the Jacobian defined by $J_{g,p}(p, \theta) = \left[\frac{\partial g_i}{\partial p_j}(p^*(\theta), \theta)\right]$ is invertible.

By the implicit function theorem, this assumption ensures that the solution to Equation (1), $p = p^*(\theta)$, is a continuously differentiable function of θ. Let θ_0 denote the true value of θ. Define

$$M(\theta) = \mathrm{E}_{\theta_0}[f(Y_i, p^*(\theta))] \qquad \text{for } \theta \in \Theta,$$

where the expectation is taken with respect to $\mathbb{P}_{\theta_0}$. The following assumptions are essentially the same as those in Section 3.2.

Assumption P5. The true value θ_0 is in the interior of Θ.

Assumption P6. $M(\theta)$ is a continuous function and has a unique maximum at θ_0 in Θ.

This assumption ensures that the true parameter θ_0 is identifiable.

Assumption P7. $f(y, p) \leq 0$ for any $(y, p) \in \mathbb{R}^{1+J}$ and $f(y, p) \in C(\mathbb{R} \times \mathbb{R}^J)$. Moreover, if Y_i is not bounded, $f(y, p)$ satisfies that for any compact set $D \subset \mathbb{R}^J$,

$$\liminf_{|y| \to \infty} \frac{1 + \inf_{x \in D}[-f(y, p)]}{\sup_{p \in D}[-f(y, p)]} > 0.$$

This assumption holds for square loss functions, i.e., $f(y,p) = -(y-p)^2$. Given any $\theta \in \Theta$ and a positive ω, recall the sieve estimate of p is given by

$$\hat{p}(\theta) = \arg\max_{p} \frac{1}{n} \sum_{i=1}^{n} f(Y_i, p) - \omega \|p - \Psi(p,\theta)\|_2^2. \tag{A12}$$

The following theorem indicates the approximate solution to the structural equation (1) is uniformly close to the exact solution $p^*(\theta)$.

Theorem A3. *Assume that Assumptions P1-P7 are satisfied. Then, we have*

$$\sup_{\theta\in\Theta} \|\hat{p}(\theta) - p^*(\theta)\|_2 = O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right), \tag{A13}$$

provided that $\omega \to \infty$ as n approaches infinity.

Proof of Theorem A3. Define $\mathbb{P}_n f(Y,p) = \frac{1}{n}\sum_{i=1}^{n} f(Y_i,p)$. To continue the proof, we first show the following technical result:

$$\sup_{\theta\in\Theta} \mathbb{P}_n[-f(Y, p^*(\theta))] = O_{\mathbb{P}}(1). \tag{A14}$$

As Assumption P2 is met, $\|p^*(\theta)\|_2 \le C$ for some positive constant C for any $\theta \in \Theta$. If Y_i's are bounded, $\sup_{\theta\in\Theta} \mathbb{P}_n[-f(Y,p^*(\theta))]$ must be bounded, because f is a continuous function. Hence, (A14) holds. If Y_i's are not bounded, by Assumption P7, there exists a positive constant η such that

$$\eta \sup_{x\in\Lambda}[-f(y,x)] \le 1 + \inf_{x\in\Lambda}[-f(y,x)] \quad \forall y \in \mathbb{R}.$$

Therefore,

$$\sup_{\theta\in\Theta} \mathbb{P}_n[-f(Y,p^*(\theta))] \le \eta^{-1}\{1 + \mathbb{P}_n[-f(Y,p^*(\theta_0))]\}.$$

By the strong law of large numbers, with probability one, we have

$$\mathbb{P}_n[-f(Y,p^*(\theta_0))] \to \mathrm{E}_{\theta_0}[-f(Y,p^*(\theta_0))] = -M(\theta_0) < \infty.$$

It follows that $\eta^{-1}\{1 + \mathbb{P}_n[-f(Y,p(\theta_0))]\} = O_{\mathbb{P}}(1)$. Equation (A14) is established.

Let $\rho(p,\Psi(p,\theta)) = \|p - \Psi(p,\theta)\|_2^2$. Based on the definition of $\hat{p}(\theta)$, we have that,

for any $\theta \in \Theta$,

$$\ell_n(\hat{p}(\theta)) - \omega\rho(\hat{p}(\theta), \Psi(\hat{p}(\theta), \theta)) \geq \ell_n(p^*(\theta)).$$

As $\ell_n(\hat{p}(\theta)) \leq 0$ by Assumption P7, it follows that

$$-\omega\rho(\hat{p}(\theta), \Psi(\hat{p}(\theta), \theta)) \geq \ell_n(p^*(\theta)).$$

Then,

$$\rho(\hat{p}(\theta), \Psi(\hat{p}(\theta), \theta)) \leq -\omega^{-1}\ell_n(p^*(\theta)) = \omega^{-1}\mathbb{P}_n[-f(Y, p^*(\theta))].$$

Therefore, it follows from (A14) that $\sup_{\theta\in\Theta} \rho(\hat{p}(\theta), \Psi(\hat{p}(\theta), \theta)) = \omega^{-1}O_{\mathbb{P}}(1)$. Consequently,

$$\sup_{\theta\in\Theta} \|g(\hat{p}(\theta), \theta)\|_2 = O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right),$$

while $g(p^*(\theta), \theta) = 0$. By Assumption P3, we have

$$\sup_{\theta\in\Theta} \|\hat{p}(\theta) - p^*(\theta)\|_2 \leq O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right).$$

This completes the proof. □

Remark A1. The error term originates from using a finite ω. Since ω is finite, the solution to the penalized optimization problem in (A11) is affected by the sample through the likelihood function ℓ_n. Therefore, there exist discrepancies between this estimator and the solution to the structural equation (1), which is the also the minimizer of the penalty term $\rho(p, \Psi(p, \theta))$ for any given θ.

To establish the consistency of the joint and nested estimators of θ, we need a stronger version of Assumption P7.

Assumption P8. $f(y, p) \leq 0$ for any $(y, p) \in \mathbb{R}^{1+J}$ and $f(y, p) \in C(\mathbb{R} \times \mathbb{R}^J)$ satisfies

$$\mathrm{E}_{\theta_0}\left[\left\|\frac{\partial f}{\partial p}(Y_i, p^*(\theta_0))\right\|_2\right] < \infty.$$

Moreover, if Y_i is not bounded, $f(y,p)$ satisfies that for any compact set $D \subset \mathbb{R}^J$,

$$\liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}[-f(y,p)]}{\sup_{p\in D}[-f(y,p)]} > 0 \quad \text{and} \quad \liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}|\partial f(y,p)/\partial p_j|}{\sup_{p\in D}|\partial f(y,p)/\partial p_j|} > 0$$

for $j = 1, \ldots, d_1$.

C.1 Asymptotic properties of the joint estimator

We first study the asymptotic properties of the joint estimator. Let $\tilde{\theta}_n$ denote the estimator of θ obtained from the joint algorithm in Section 3.1. Actually, $\tilde{\theta}_n$ is defined by

$$\tilde{\theta}_n = \arg\max_{\theta\in\Theta} \ell_n(\hat{p}(\theta)) - w\|\hat{p}(\theta) - \Psi(\hat{p}(\theta),\theta)\|_2^2, \tag{A15}$$

where $\hat{p}(\theta)$ is given by (A12).

Theorem A4. *Suppose that Assumptions P1-P6 and P8 hold. If $\omega \to \infty$, then $\tilde{\theta}_n$ is consistent.*

Proof of Theorem A4. We first show that

$$\sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right) + o_{\mathbb{P}}\left(\frac{1}{n}\right). \tag{A16}$$

Recall that $M(\theta) = \mathrm{E}_{\theta_0}[f(Y_i, p^*(\theta))]$. Then, we define

$$M_n(\theta) = \ell_n(p^*(\theta)) = \frac{1}{n}\sum_{i=1}^{n} f(Y_i, p^*(\theta)) \quad \text{for } \theta \in \Theta.$$

We will show

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| = o_{\mathbb{P}}(1). \tag{A17}$$

Finally, we prove that

$$\tilde{\theta}_n \to \theta_0$$

in probability as $n \to \infty$.

To prove (A16), we assume that Θ is convex without loss of generality. By Assumption P2, there must exist some positive constant r such that

$$\|p^*(\theta)\|_2 \le r, \qquad \forall \theta \in \Theta.$$

Let $V_n = \sup_{\theta\in\Theta} \|\hat{p}(\theta) - p^*(\theta)\|_2$. Theorem A3 indicates

$$V_n \leq O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right).$$

By Assumption P8, there exists a positive constant η such that, for $j = 1, \ldots, J$,

$$\sup_{\|p\|_2\leq r+1} \left|\frac{\partial f}{\partial p_j}(y,p)\right| \leq \frac{1}{\eta}\left\{1 + \inf_{\|p\|_2\leq r+1}\left|\frac{\partial f}{\partial p_j}(y,p)\right|\right\} \quad \forall y \in \mathbb{R}. \tag{A18}$$

Then, it follows that

$$\begin{aligned}
&|\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| \\
\leq& \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))| \\
=& \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))|\mathbb{1}_{(V_n\leq 1)} + \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))|\mathbb{1}_{(V_n>1)} \\
\leq& \frac{1}{n}\sum_{i=1}^n \left[\sup_{\substack{\|p\|_2\leq r+1,\\ 1\leq j\leq J}} \left|\frac{\partial f}{\partial p_j}(Y_i,p)\right|\right] \|\hat{p}(\theta) - p^*(\theta)\|_2 \mathbb{1}_{(V_n\leq 1)} \\
&\qquad + \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))|\mathbb{1}_{(V_n>1)} \\
\leq& \frac{1}{n}\sum_{i=1}^n \frac{1}{\eta}\left\{1 + \max_{1\leq j\leq J}\inf_{\|p\|_2\leq r+1}\left|\frac{\partial f}{\partial p_j}(Y_i,p)\right|\right\} V_n \mathbb{1}_{(V_n\leq 1)} \\
&\qquad + \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))|\mathbb{1}_{(V_n>1)} \\
\leq& \frac{1}{n}\sum_{i=1}^n \frac{1}{\eta}\left\{1 + \left\|\frac{\partial f}{\partial p}(Y_i, p^*(\theta_0))\right\|_2\right\} V_n + \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))|\mathbb{1}_{(V_n>1)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| \leq& \frac{1}{n}\sum_{i=1}^n \frac{1}{\eta}\left\{1 + \left\|\frac{\partial f}{\partial p}(Y_i, p^*(\theta_0))\right\|_2\right\} V_n \\
&+ \frac{1}{n}\sum_{i=1}^n |f(Y_i, \hat{p}(\theta)) - f(Y_i, p^*(\theta))|\mathbb{1}_{(V_n>1)}.
\end{aligned} \tag{A19}$$

By the strong law of large numbers and Assumption P8,

$$\frac{1}{n}\sum_{i=1}^{n}\left[1+\left\|\frac{\partial f}{\partial p}(Y_i,p^*(\theta_0))\right\|_2\right]\to 1+\mathrm{E}_{\theta_0}\left\|\frac{\partial f}{\partial p}(Y_i,p^*(\theta_0))\right\|_2$$

almost surely. So, it is $O_{\mathbb{P}}(1)$. The second term on the right-hand side of (A19) is nonzero only in the event $\{V_n>1\}$, whose probability converges to zero by Theorem A3, so it is $o_{\mathbb{P}}(n^{-1})$. Hence, we have established Equation (A16). Equation (A17) follows from Lemma A6, which will be presented later.

Now, we are ready to prove $\tilde{\theta}_n\to\theta_0$ in probability. Let δ be an arbitrary positive number. Assumption P6 indicates that θ_0 is the unique maximizer of $M(\theta)$. As $M(\theta)$ is continuous over the compact set Θ,

$$\gamma:=M(\theta_0)-\sup_{\theta\in\Theta,d(\theta,\theta_0)\geq\delta}M(\theta)>0, \tag{A20}$$

where $d(\theta,\theta_0)=\|\theta-\theta_0\|_2$ for any $\theta\in\Theta$. Note that $\tilde{\theta}_n\in\Theta$. By (A16) and (A17), we have

$$\begin{aligned}
&M(\theta_0)-M(\tilde{\theta}_n)\\
&=M_n(\theta_0)-M_n(\tilde{\theta}_n)+o_{\mathbb{P}}(1)\\
&=\ell_n(p^*(\theta_0))-\ell_n(p^*(\tilde{\theta}_n))+o_{\mathbb{P}}(1)\\
&\leq\ell_n(p^*(\theta_0))-\ell_n(\hat{p}(\tilde{\theta}_n))+\sup_{\theta\in\Theta}|\ell_n(\hat{p}(\theta))-\ell_n(p^*(\theta))|+o_{\mathbb{P}}(1)\\
&=\ell_n(p^*(\theta_0))-\ell_n(\hat{p}(\tilde{\theta}_n))+o_{\mathbb{P}}(1).
\end{aligned}$$

By definition of $\tilde{\theta}_n$ and $p^*(\theta_0)$, we have

$$\ell_n(\hat{p}(\tilde{\theta}_n))\geq\ell_n(p^*(\theta_0)),$$

so $M(\theta_0)-M(\tilde{\theta}_n)\leq o_{\mathbb{P}}(1)$. By (A20), one has

$$\{d(\tilde{\theta}_n,\theta_0)\geq\delta\}\subset\{M(\theta_0)-M(\tilde{\theta}_n)\geq\gamma\}\subset\{o_{\mathbb{P}}(1)\geq\gamma\}.$$

Therefore,

$$\mathbb{P}_{\theta_0}(d(\tilde{\theta}_n,\theta_0)\geq\delta)\leq\mathbb{P}_{\theta_0}(o_{\mathbb{P}}(1)\geq\gamma),$$

which converges to 0 as $n \to \infty$. As δ is an arbitrary positive number, $\tilde{\theta}_n$ is a consistent estimator of θ_0. This completes the proof. □

Remark A2. Since $\tilde{\theta}_n$ is the maximizer of $\ell_n(\hat{p}(\theta))$ with a non-positive penalty, we choose $\ell_n(\hat{p}(\theta))$ as the criterion function. Though it is difficult to evaluate the gradient of $\hat{p}(\theta)$ with respect to θ, we can still use it as our criterion function, because we have established the bound on the difference between $\hat{p}(\theta)$ and $p^*(\theta)$ in Theorem A3. We establish the consistency and later the asymptotic normality of $\tilde{\theta}_n$ by resorting to the techniques for M-estimators. Even though $\tilde{\theta}_n$ is not the maximizer of $\ell_n(p^*(\theta))$, we are able to control the difference between $\ell_n(p^*(\tilde{\theta}_n))$ and $\max_{\theta\in\Theta} \ell_n(p^*(\theta))$. We show that, in actuality, $\tilde{\theta}_n$ nearly maximizes $\ell_n(p^*(\theta))$, and then we leverage the results for M-estimators; see Section 5.2 of Van der Vaart and Wellner (1996) for more details.

To derive asymptotic normality of $\tilde{\theta}_n$, we need a stronger condition than Assumptions P7 and P8.

Assumption P9. $f(y,p) \leq 0$ for any $(y,p) \in \mathbb{R}^{1+J}$ and $f(y,p) \in C(\mathbb{R}\times\mathbb{R}^J)$ satisfies

$$\mathrm{E}_{\theta_0}\left[\left\|\frac{\partial f}{\partial p}(Y,p^*(\theta_0))\right\|_2^2\right] < \infty \quad \text{and} \quad \mathrm{E}_{\theta_0}\left[\left|\frac{\partial^2 f}{\partial p_i \partial p_j}(Y,p^*(\theta_0))\right|\right] < \infty$$

for $i,j = 1,\ldots,J$. Moreover, if Y_i is not bounded, $f(y,p)$ satisfies that for any compact set $D \subset \mathbb{R}^J$,

$$\liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}[-f(y,p)]}{\sup_{p\in D}[-f(y,p)]} > 0 \qquad \liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}|\partial f(y,p)/\partial p_j|}{\sup_{p\in D}|\partial f(y,p)/\partial p_j|} > 0$$

for $j = 1,\ldots,J$, and

$$\liminf_{|y|\to\infty} \frac{1+\inf_{p\in D}|\partial^2 f(y,p)/\partial p_i \partial p_j|}{\sup_{p\in D}|\partial^2 f(y,p)/\partial p_i \partial p_j|} > 0$$

for any $i,j = 1,\ldots,J$.

Theorem A5. *Suppose that Assumptions P1-P6 and P9 hold. If $\omega/n^2 \to \infty$ and the matrix*

$$V_{\theta_0} = \mathrm{E}_{\theta_0}\left[\left\{\frac{\partial p^*(\theta_0)}{\partial \theta}\right\}' \frac{\partial^2 f(Y_i,p^*(\theta_0))}{\partial p \partial p'} \left\{\frac{\partial p^*(\theta_0)}{\partial \theta}\right\}\right]$$

is invertible, then

$$\sqrt{n}\,(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

where $\Sigma = V_{\theta_0}^{-1} \left(\frac{\partial p^*(\theta_0)}{\partial \theta} \right)' \mathrm{E}_{\theta_0} \left[\left\{ \frac{\partial f(y,p^*(\theta_0))}{\partial p} \right\} \left\{ \frac{\partial f(y,p^*(\theta_0))}{\partial p} \right\}' \right] \left(\frac{\partial p^*(\theta_0)}{\partial \theta} \right) V_{\theta_0}^{-1}$.

Remark A3. Under mild regularity conditions on f,

$$-\mathrm{E}_{\theta_0} \left[\frac{\partial^2 f(Y_i, p^*(\theta_0))}{\partial p \partial p'} \right] = \mathrm{E}_{\theta_0} \left[\left\{ \frac{\partial f(y, p^*(\theta_0))}{\partial p} \right\} \left\{ \frac{\partial f(y, p^*(\theta_0))}{\partial p} \right\}' \right],$$

then $\Sigma = -V_{\theta_0}^{-1}$. Consequently, even though $\tilde{\theta}_n$ is different from the maximum likelihood estimator of θ, which minimizes $\ell_n(p^*(\theta))$, the asymptotic variance of $\hat{\theta}_n$ attains the Cramér-Rao lower bound. Thus, $\tilde{\theta}_n$ is asymptotically efficient.

Proof of Theorem A5. We mainly follow Theorem 5.23 of Van der Vaart (2000) to prove asymptotic normality of $\tilde{\theta}_n$. Firstly, as we have shown in the proof of Lemma A6,

$$\begin{aligned} & |f(y, p^*(\theta_1)) - f(y, p^*(\theta_2))| \\ \leq & \frac{1}{\eta} \left[1 + \left\| \frac{\partial f}{\partial p}(y, p^*(\theta_0)) \right\|_2 \right] \|p^*(\theta_1) - p^*(\theta_2)\|_2 \\ \leq & \frac{C}{\eta} \left[1 + \left\| \frac{\partial f}{\partial p}(y, p^*(\theta_0)) \right\|_2 \right] \|\theta_1 - \theta_2\|_2 \end{aligned} \tag{A21}$$

for some constant C, and η is defined in Equation (A18). By Assumption P9, the right-hand side of (A21) has a finite second moment.

Next, we consider a second-order Tayor expansion for

$$M(\theta) = \mathrm{E}_{\theta_0}[f(Y, p^*(\theta))]$$

in a neighbourhood of θ_0. Obviously,

$$\begin{aligned} & f(y, p(\theta)) \\ = & f(y, p^*(\theta_0)) \\ & + \left[\frac{\partial f(y, p^*(\theta_0))}{\partial p} \right]' \left(\frac{\partial p^*(\theta_0)}{\partial \theta} \right) (\theta - \theta_0) \\ & + \frac{1}{2} (\theta - \theta_0)' \left[\sum_{j=1}^{J} \frac{\partial f(y, p^*(\theta_0))}{\partial p_j} \frac{\partial^2 p_j^*(\theta_0)}{\partial \theta \partial \theta'} + \left(\frac{\partial p^*(\theta_0)}{\partial \theta} \right)' \frac{\partial^2 f(y, p^*(\theta_0))}{\partial p \partial p'} \left(\frac{\partial p^*(\theta_0)}{\partial \theta} \right) \right] \\ & \times (\theta - \theta_0) + R, \end{aligned} \tag{A22}$$

where R is the remainder term. Define

$$D(y,\theta) = \sum_{j=1}^{J} \frac{\partial f(y,p^*(\theta))}{\partial p_j} \frac{\partial^2 p_j^*(\theta)}{\partial\theta\partial\theta'} + \left(\frac{\partial p^*(\theta)}{\partial\theta}\right)' \frac{\partial^2 f(y,p^*(\theta))}{\partial p\partial p'} \left(\frac{\partial p^*(\theta)}{\partial\theta}\right).$$

Then the reminder term can be rewritten as

$$R = (\theta-\theta_0)' \left[\int_0^1 [D(y,\theta_0+s(\theta-\theta_0)) - D(y,\theta_0)](1-s)\,\mathrm{d}s\right](\theta-\theta_0).$$

Note that $D(y,\theta)$ is a $p\times p$ matrix. For any (a,b)th entry in $D(y,\theta)$, by using the same argument for Equation (A19), we can show that, for any $\theta\in\Theta$,

$$\begin{aligned}
&|D_{ab}(y,\theta)| \\
&\le \sum_{j=1}^{J} \left|\frac{\partial f(y,p^*(\theta))}{\partial p_j}\frac{\partial^2 p_j^*(\theta)}{\partial\theta_a\partial\theta_b}\right| + \left|\left(\frac{\partial p^*(\theta)}{\partial\theta_a}\right)' \frac{\partial^2 f(y,p^*(\theta))}{\partial p\partial p'} \left(\frac{\partial p^*(\theta)}{\partial\theta_b}\right)\right| \\
&\le \frac{C'}{\eta}\left[1 + \left\|\frac{\partial f}{\partial p}(y,p^*(\theta_0))\right\|_2 + \max_{i,j}\left|\frac{\partial^2 f}{\partial p_i\partial p_j}(y,p^*(\theta_0))\right|\right],
\end{aligned}$$

where C' is a positive constant. Additionally, under Assumption P9, the right-hand side of the above inequality has a finite mean. Therefore, applying the dominated convergence theorem, we have

$$\mathrm{E}_{\theta_0}\left[\int_0^1 [D(y,\theta_0+s(\theta-\theta_0)) - D(y,\theta_0)](1-s)\,\mathrm{d}s\right] \to 0$$

as $\theta\to\theta_0$. Then, by the Taylor expansion of $f(y,p^*(\theta))$, it follows that

$$M(\theta) = M(\theta_0) + \frac{1}{2}(\theta-\theta_0)' V_{\theta_0}(\theta-\theta_0) + o\left(\|\theta-\theta_0\|_2^2\right). \tag{A23}$$

Recall that $f(y,p)$ is the log density of Y_i. Therefore, there is no linear form in (A23) and the expected value of $D(Y,\theta_0)$ is given by V_{θ_0} as the expectation of (A22) is zero.

Finally, we want to establish that

$$\ell_n(p^*(\tilde{\theta}_n)) \ge \sup_{\theta\in\Theta} \ell_n(p^*(\theta)) - o_{\mathbb{P}}(n^{-1}). \tag{A24}$$

By definition of $p^*(\theta)$ and $\tilde{\theta}_n$, we have

$$\begin{aligned}
&\ell_n(p^*(\tilde{\theta}_n)) \\
&\geq \ell_n(\hat{p}(\tilde{\theta}_n)) - \sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| \\
&\geq \sup_{\theta\in\Theta} \ell_n(p^*(\theta)) - \sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))|.
\end{aligned}$$

By (A16) and $\omega/n^2 \to \infty$, we have

$$\sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right) + o_{\mathbb{P}}\left(\frac{1}{n}\right) = o_{\mathbb{P}}(n^{-1}).$$

Hence, the relation in (A24) holds. By (A21), (A23), (A24), and Theorem A4, it follows from Theorem 5.23 in Van der Vaart (2000) that $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1} \left(\frac{\partial p^*(\theta_0)}{\partial \theta}\right)' \mathrm{E}_{\theta_0}\left[\left\{\frac{\partial f(y, p^*(\theta_0))}{\partial p}\right\}\left\{\frac{\partial f(y, p^*(\theta_0))}{\partial p}\right\}'\right] \left(\frac{\partial p^*(\theta_0)}{\partial \theta}\right) V_{\theta_0}^{-1},$$

if V_{θ_0} is non-singular. This completes the proof. □

Lemma A6. *The class* $\{f(\cdot, p^*(\theta)), \theta \in \Theta\}$ *is* $\mathbb{P}_{\theta_0}$*-Glivenko-Cantelli.*

Proof. As Assumption P4 is met and Θ is a compact set, by the implicit function theorem, there exists some constant C such that

$$\|p^*(\theta_1) - p^*(\theta_2)\|_2 \leq C\|\theta_1 - \theta_2\|_2$$

for any $\theta_1, \theta_2 \in \Theta$. Furthermore, with a similar argument for Equation (A19), we obtain

$$\begin{aligned}
&|f(y, p^*(\theta_1)) - f(y, p^*(\theta_2))| \\
&\leq \frac{1}{\eta}\left[1 + \left\|\frac{\partial f}{\partial p}(y, p^*(\theta_0))\right\|_2\right] \|p^*(\theta_1) - p^*(\theta_2)\|_2 \\
&\leq \frac{C}{\eta}\left[1 + \left\|\frac{\partial f}{\partial p}(y, p^*(\theta_0))\right\|_2\right] \|\theta_1 - \theta_2\|_2. \qquad \text{(A25)}
\end{aligned}$$

By Assumption P8, $\left\|\frac{\partial f}{\partial p}(Y, p^*(\theta_0))\right\|_2$ has a finite expectation under $\mathbb{P}_{\theta_0}$. Thus, based on Theorem 2.7.11 in Van der Vaart and Wellner (1996), the $L_1(\mathbb{P}_{\theta_0})$-bracketing

number is bounded by the covering number $N(\epsilon, \Theta_0, \|\cdot\|_2)$ of Θ_0. Since Θ_0 is a compact subset of $\mathbb{R}^d$,

$$N(\epsilon, \Theta, \|\cdot\|) \leq C_0 \times \left(\frac{1}{\epsilon}\right)^d$$

for some constant C_0 and any $\epsilon > 0$. Therefore, by Theorem 2.4.1 of Van der Vaart and Wellner (1996), this lemma holds. □

C.2 Asymptotic Properties for the Nested Estimator

Recall that $\hat{\theta}_n$ obtained from the nested algorithm is the maximizer of $\ell_n(\hat{p}(\theta))$. Similar to the joint estimator, we now establish consistency for $\hat{\theta}_n$ as an estimator of θ_0 by invoking M-estimation techniques.

Theorem A7. *Under conditions of Theorem A4, $\hat{\theta}_n$ is consistent.*

Proof of Theorem A7. We employ the techniques of proving Theorem A4. By (A16) and (A17), we have

$$\begin{aligned}
& M(\theta_0) - M(\hat{\theta}_n) \\
&= M_n(\theta_0) - M_n(\hat{\theta}_n) + o_{\mathbb{P}}(1) \quad \text{by (A17)} \\
&= \ell_n(p^*(\theta_0)) - \ell_n(p^*(\hat{\theta}_n)) + o_{\mathbb{P}}(1) \\
&\leq \ell_n(\hat{p}(\theta_0)) + \sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| - \ell_n(\hat{p}(\hat{\theta}_n)) \\
&\quad + \sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| + o_{\mathbb{P}}(1) \\
&= \ell_n(\hat{p}(\theta_0)) - \ell_n(\hat{p}(\hat{\theta}_n)) + 2\sup_{\theta\in\Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| + o_{\mathbb{P}}(1) \\
&\leq \ell_n(\hat{p}(\theta_0)) - \ell_n(\hat{p}(\hat{\theta}_n)) + o_{\mathbb{P}}(1).
\end{aligned}$$

By definition of $\hat{\theta}_n$, we have

$$\ell_n(\hat{p}(\hat{\theta}_n)) \geq \ell_n(\hat{p}(\theta_0)),$$

so $M(\theta_0) - M(\hat{\theta}_n) \leq o_{\mathbb{P}}(1)$. By (A20), we have

$$\{d(\hat{\theta}_n, \theta_0) \geq \delta\} \subset \{M(\theta_0) - M(\hat{\theta}_n) \geq \gamma\} \subset \{o_{\mathbb{P}}(1) \geq \gamma\}.$$

Therefore,

$$\mathbb{P}_{\theta_0}(d(\hat{\theta}_n, \theta_0) \geq \delta) \leq \mathbb{P}_{\theta_0}(o_{\mathbb{P}}(1) \geq \gamma),$$

which converges to 0 as n approaches infinity. As δ is an arbitrary positive number, $\hat{\theta}_n$ is a consistent estimator of θ_0. This completes the proof. □

The following theorem indicates the nested estimator has the same asymptotic distribution as the joint estimator under mild conditions.

Theorem A8. *Under conditions of Theorem A5,*

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Proof of Theorem A8. We leverage the same techniques of proving Theorem A5. In other words, we only need to show

$$\ell_n(p^*(\hat{\theta}_n)) \geq \sup_{\theta \in \Theta} \ell_n(p^*(\theta)) - o_{\mathbb{P}}(n^{-1}). \tag{A26}$$

Note that

$$\begin{aligned}
&\ell_n(p^*(\hat{\theta}_n)) \\
\geq\ & \ell_n(\hat{p}(\hat{\theta}_n)) - \sup_{\theta \in \Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| \\
=\ & \sup_{\theta \in \Theta} \ell_n(\hat{p}(\theta)) - \sup_{\theta \in \Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| \\
\geq\ & \sup_{\theta \in \Theta} [\ell_n(p^*(\theta)) - \sup_{\theta \in \Theta} |\ell_n(p^*(\theta)) - \ell_n(\hat{p}(\theta))|] - \sup_{\theta \in \Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| \\
\geq\ & \sup_{\theta \in \Theta} \ell_n(p^*(\theta)) - 2 \sup_{\theta \in \Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))|.
\end{aligned}$$

By (A16) and $\omega/n^2 \to \infty$, we have

$$\sup_{\theta \in \Theta} |\ell_n(\hat{p}(\theta)) - \ell_n(p^*(\theta))| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{\omega}}\right) + o_{\mathbb{P}}\left(\frac{1}{n}\right) = o_{\mathbb{P}}(n^{-1}).$$

Hence, the relation in (A26) is verified. By (A21), (A23), (A26), and Theorem A7, it follows from Theorem 5.23 in Van der Vaart (2000) that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically

normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1}\left(\frac{\partial p^*(\theta_0)}{\partial \theta}\right)' \mathrm{E}_{\theta_0}\left[\left\{\frac{\partial f(y,p^*(\theta_0))}{\partial p}\right\}\left\{\frac{\partial f(y,p^*(\theta_0))}{\partial p}\right\}'\right]\left(\frac{\partial p^*(\theta_0)}{\partial \theta}\right)V_{\theta_0}^{-1},$$

if V_{θ_0} is non-singular. This completes the proof. □

D More about the Joint and Nested algorithms

In this section, we develop more properties on these two algorithms.

The joint algorithm is attractive because it involves a single-level optimization problem and computes the Hessian matrix with respect to (β, θ) at the solution directly. The following corollary provides a natural way to calculate the standard error of $\tilde{\theta}_n$ using the Hessian matrix generated from the joint algorithm.

Corollary A8.1. *The Fisher information can be characterized as*

$$\widehat{H} = \mathbf{H}_{\theta\theta} - \mathbf{H}'_{\beta\theta}\mathbf{H}^{-1}_{\beta\beta}\mathbf{H}_{\beta\theta},$$

where the matrices in bold are the four blocks in the Hessian matrix generated from a joint maximization algorithm,

$$\begin{bmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\theta} \\ \mathbf{H}_{\theta\beta} & \mathbf{H}_{\theta\theta} \end{bmatrix}.$$

Proof of Corollary A8.1. Consider the following problem: $\max \quad h(\beta,\theta) = \ell(\beta,\theta) - \omega\left(\Psi(\beta,\theta) - \beta\right)^2$ where $\beta = p$. Taking the first-order condition gives

$$\begin{aligned} \ell_\beta - 2\omega\left(\Psi - \beta\right)\left(\Psi_\beta - 1\right) &= 0 \\ \ell_\theta - 2\omega\left(\Psi - \beta\right)\Psi_\theta &= 0 \end{aligned}$$

Note that $\Psi(\widehat{\beta}(\theta),\theta) - \widehat{\beta}(\theta) \approx 0$ for ω approaches infinity, which further implies that $\ell_\beta \approx 0$. Taking the derivative gives $\Psi_\beta\widehat{\beta}'(\theta) + \Psi_\theta - \widehat{\beta}'(\theta) \approx 0$, which implies that $\Psi_\beta - 1 \approx -\frac{\Psi_\theta}{\widehat{\beta}'(\theta)}$.

Taking the second-order derivative gives $\begin{pmatrix} \frac{\partial^2 h}{\partial\theta\partial\theta} & \frac{\partial^2 h}{\partial\theta\partial\beta} \\ \frac{\partial^2 h}{\partial\theta\partial\beta} & \frac{\partial^2 h}{\partial\beta\partial\beta} \end{pmatrix}$, where

$$\begin{aligned}
\frac{\partial^2 h}{\partial\theta\partial\theta} &= \ell_{\theta\theta} - 2\omega\left[(\Psi_\theta)^2 + (\Psi-\beta)\,\Psi_{\theta\theta}\right] \\
\frac{\partial^2 h}{\partial\theta\partial\beta} &= \ell_{\theta\beta} - 2\omega\left[(\Psi_\beta - 1)\,\Psi_\theta + (\Psi-\beta)\,\Psi_{\beta\theta}\right] \\
\frac{\partial^2 h}{\partial\beta\partial\beta} &= \ell_{\beta\beta} - 2\omega\left[(\Psi_\beta - 1)^2 + (\Psi-\beta)\,\Psi_{\beta\beta}\right]
\end{aligned}$$

As ω approaches infinity, we study the block that we highlight here[15]

$$\begin{aligned}
&\mathbf{H}_{\theta\theta} - \mathbf{H}'_{\beta\theta}\mathbf{H}^{-1}_{\beta\beta}\mathbf{H}_{\beta\theta} \\
=&\ell_{\theta\theta} - 2\omega\,(\Psi_\theta)^2 - \frac{\left[\ell_{\theta\beta} + 2\omega\frac{(\Psi_\theta)^2}{\widehat{\beta}'(\theta)}\right]^2}{\ell_{\beta\beta} - 2\omega\left(\frac{\Psi_\theta}{\widehat{\beta}'(\theta)}\right)^2} \\
=&\ell_{\theta\theta} - \frac{2\omega\,(\Psi_\theta)^2\,\ell_{\beta\beta} - 4\omega^2\,(\Psi_\theta)^2\left(\frac{\Psi_\theta}{\widehat{\beta}'(\theta)}\right)^2 + (\ell_{\theta\beta})^2 + 4\omega^2\frac{(\Psi_\theta)^4}{(\widehat{\beta}'(\theta))^2} + 4\omega\ell_{\theta\beta}\frac{(\Psi_\theta)^2}{\widehat{\beta}'(\theta)}}{\ell_{\beta\beta} - 2\omega\left(\frac{\Psi_\theta}{\widehat{\beta}'(\theta)}\right)^2} \\
\to&\ell_{\theta\theta} + \ell_{\beta\beta}\left[\widehat{\beta}'(\theta)\right]^2 + 2\ell_{\theta\beta}\widehat{\beta}'(\theta)
\end{aligned}$$

Now consider $\ell(\widehat{\beta}(\theta),\theta)$, where $\widehat{\beta}(\theta)$ solves $\Psi(\beta,\theta)=\beta$. We have the Hessian

$$\ell_{\theta\theta} + \ell_{\beta\beta}\left[\widehat{\beta}'(\theta)\right]^2 + 2\ell_{\theta\beta}\widehat{\beta}'(\theta) + \ell_\beta\widehat{\beta}''(\theta) = \ell_{\theta\theta} + \ell_{\beta\beta}\left[\widehat{\beta}'(\theta)\right]^2 + 2\ell_{\theta\beta}\widehat{\beta}'(\theta),$$

which equals the limit of $\mathbf{H}_{\theta\theta} - \mathbf{H}'_{\beta\theta}\mathbf{H}^{-1}_{\beta\beta}\mathbf{H}_{\beta\theta}$. □

Next we derive the gradient of the outer loop problem in the nested algorithm.

Proposition A1. *The gradient of the outer loop satisfies*

$$\nabla\widehat{\ell}(\theta) = \nabla_\theta\ell(\widehat{\beta}(\theta),\theta) + (\nabla\widehat{\beta}(\theta))' \times [\nabla_\beta\ell(\widehat{\beta}(\theta),\theta)].$$

Proof of Proposition A1. Fix the smoothing parameter ω for the moment. Recall

[15]Note that the inverse of a block matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A-BD^{-1}C)^{-1} & * \\ * & * \end{pmatrix}$.

that $\widehat{\beta}(\theta)$ is defined implicitly by an equation system

$$\frac{\partial \ell(\beta,\theta)}{\partial \beta_k} - \omega \frac{\partial \rho(\beta,\theta)}{\partial \beta_k} = 0.$$

Denote this system as $h_k(\beta,\theta) = 0$, where $k = 1,\ldots,K$. Taking the derivative (w.r.t. θ) gives $\sum_\ell \frac{\partial h_k}{\partial \beta_\ell}\frac{d\beta_\ell}{d\theta} + \frac{\partial h_k}{\partial \theta} = 0$. It allows us to find $\nabla\widehat{\beta}(\theta)$ using the Hessian of h with respect to β, i.e., $\mathbf{H}_{\beta\beta}(\theta)$, and the cross derivative of h with respect to β and θ, i.e. $\mathbf{H}_{\beta\theta}(\theta)$, as follows:

$$\nabla\widehat{\beta}(\theta) = -\mathbf{H}_{\beta\beta}(\theta)^{-1}\mathbf{H}_{\beta\theta}(\theta),$$

where the terms on the RHS are calculated at the inner loop solution $\beta = \widehat{\beta}(\theta)$. Therefore, the gradient of $\ell(p^{\widehat{\beta}(\theta)})$ w.r.t. θ can be calculated as follows

$$\frac{\partial \widehat{\ell}(\theta)}{\partial \theta_\ell} = \left.\frac{\partial \ell(\beta,\theta)}{\partial \theta_\ell}\right|_{\beta=\widehat{\beta}(\theta)} + \sum_{k=1}^{K} \left.\frac{\partial \ell(\beta,\theta)}{\partial \beta_k}\right|_{\beta=\widehat{\beta}(\theta)} \frac{\partial \widehat{\beta}_k(\theta)}{\partial \theta_\ell},$$

which can be written in matrix form. □

E Additional Simulations

The DGP is only for demonstration purposes in the Monte Carlo simulations in the main text. We now conduct additional simulation experiments in a rich setting building on our empirical application to further demonstrate our method. Specifically, we constructed a richer DGP using market- and player-specific variables along with the maximum likelihood estimates. We set the true parameters to the rounded estimates. We sampled markets (with replacement) using the same dataset and re-generated Walmart and Kmart's choices using the equilibrium CCPs. Table E1 presents the simulation results based on 100 replications with sample sizes of $n = 2000$ and $n = 4000$, respectively. All simulations were implemented in Matlab R2024b, using `fminunc` with its built-in quasi-Newton algorithm, and were run on a machine with an 11th Gen Intel Core i7-11800H 2.30GHz processor and 16GB of RAM.

Table E1: Simulation Results

		n=2000		n=4000		n=2000		n=4000	
	θ	Ours	std	Ours	std	MLE	std	MLE	std
Market-specific									
pop	3	3.0274	0.1280	2.9987	0.0968	3.0309	0.1301	3.0018	0.0933
spc	3	3.0117	0.1924	3.0092	0.1479	3.0217	0.1913	3.0153	0.1448
urban	2	2.0218	0.2914	1.9918	0.2130	2.0262	0.2962	1.9953	0.2104
Walmart-specific									
intercept	-22	-22.2541	1.7524	-22.1481	1.4207	-22.3295	1.7184	-22.1859	1.3371
dbenton	-2	-1.9875	0.1461	-1.9854	0.1039	-1.9907	0.1407	-1.9896	0.0974
south	1	1.0008	0.1560	0.9820	0.1100	1.0067	0.1473	0.9868	0.1018
Kmart-specific									
intercept_K	-36	-36.1569	1.6444	-36.0579	1.3321	-36.2479	1.6475	-36.1200	1.3150
midwest	1	1.0073	0.1452	0.9992	0.0989	1.0006	0.1437	0.9957	0.0965
Δ	2	2.0605	0.2954	2.0178	0.2337	2.0620	0.2441	2.0133	0.1802
Computation Time (seconds)		34.80	9.71	52.28	12.45	15.12	0.92	25.77	3.06

Note: We estimate the model using the proposed method when $K = 10, 20, 30$ and obtain very similar results. We report only the results for $K = 10$ in this table.

F Deriving Gradient and Hessian Functions

F.1 The Monopoly Pricing Problem

In this subsection, we derive the gradient of the objective function in the monopoly pricing problem.

In the inner loop, we maximize the following function with respect to β

$$h(\beta, \theta) = \sum_{i=1}^{n} \log \phi(y_i - p^\beta(x_i)) - \omega \sum_{\ell=1}^{L} \left[p^\beta(x_\ell) e^{p^\beta(x_\ell)} - \theta x_\ell \right]^2,$$

where $L = 1000$ is the number of grid points to approximate the integration. Note that the standard normal density $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$ and $p^\beta(x) = \sum_{k=1}^{K} \beta_k s_k(x)$. We will consider the two terms in sequence.

In the first step,

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_k} =& \sum (y_i - p^\beta(x_i)) s_k(x_i), \\ \frac{\partial^2 \ell}{\partial \beta_k \partial \beta_{k'}} =& - \sum s_k(x_i) s_{k'}(x_i). \end{aligned}$$

In the second step,

$$\begin{aligned} \frac{\partial \rho}{\partial \beta_k} =& 2 \sum_{\ell=1}^{L} [p^\beta(x_\ell) e^{p^\beta(x_\ell)} - \theta x_\ell] e^{p^\beta(x_\ell)} [1 + p^\beta(x_\ell)] s_k(x_\ell), \\ \frac{\partial^2 \rho}{\partial \beta_k \partial \beta_{k'}} =& 2 \sum_{\ell=1}^{L} \Big[\left(e^{p^\beta(x_\ell)} [1 + p^\beta(x_\ell)] \right)^2 \\ & + [p^\beta(x_\ell) e^{p^\beta(x_\ell)} - \theta x_\ell] e^{p^\beta(x_\ell)} [2 + p^\beta(x_\ell)] \Big] s_k(x_\ell) s_{k'}(x_\ell), \\ \frac{\partial^2 \rho}{\partial \beta_k \partial \theta} =& -2 \sum_{\ell=1}^{L} x_\ell e^{p^\beta(x_\ell)} [1 + p^\beta(x_\ell)] s_k(x_\ell). \end{aligned}$$

F.2 Static Game with Incomplete Information

In this subsection, we derive the gradient of the objective function in the static game with incomplete information, where

$$\ell(\beta,\theta)=\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\Big[d_{jm}\log\Big(p_j^{\beta}(\xi_m(\theta))\Big)+(1-d_{jm})\log\Big(1-p_j^{\beta}(\xi_m(\theta))\Big)\Big] \tag{A27}$$

$$\rho(\beta,\theta)=\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\Big[p_j^{\beta}(\xi_m(\theta))-\sigma\Big(\xi_{jm}-p_{-j}^{\beta}(\xi_m(\theta))\Delta\Big)\Big]^2 \tag{A28}$$

We approximate the CCP by $p_j^{\beta}(\xi_j,\xi_{-j})=\sigma\Big(\sum_{\imath=1}^{K}\sum_{\jmath=1}^{K}\beta_{\imath\jmath}s_\imath\big(\sigma(\xi_j)\big)s_\jmath\big(\sigma(\xi_{-j})\big)\Big)$. Denote $v_j=\sum_{\imath=1}^{K}\sum_{\jmath=1}^{K}\beta_{\imath\jmath}s_\imath\big(\sigma(\xi_j)\big)s_\jmath\big(\sigma(\xi_{-j})\big)$ for convenience. Note that $\sigma(v)=\frac{1}{1+e^{-v}}$ and $\sigma'(v)=\frac{-1}{\sigma^2}e^{-v}(-1)=\frac{e^{-v}}{\sigma(v)^2}$. Moreover, $\xi_{jm}=X'_{jm}\alpha-Z'_m\gamma$. We stack the parameters $\theta=(\gamma,\alpha_w,\alpha_k,\Delta)'$, where α_w and α_k are WalMart and Kmart's coefficients.

First, denote $p_{jm}=p_j^{\beta}(\xi_m(\theta))$.

$$\begin{aligned}
\frac{\partial\ell}{\partial\beta_{\imath\jmath}}=&\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\left[\frac{d_{jm}}{p_{jm}}-\frac{1-d_{jm}}{1-p_{jm}}\right]\cdot\frac{\partial p_{jm}}{\partial\beta_{\imath\jmath}},\\
\frac{\partial\ell}{\partial\theta_k}=&\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\left[\frac{d_{jm}}{p_{jm}}-\frac{1-d_{jm}}{1-p_{jm}}\right]\cdot\frac{\partial p_{jm}}{\partial\theta_k},\\
\frac{\partial\ell}{\partial\Delta}=&0,\\
\frac{\partial\rho}{\partial\beta_{\imath\jmath}}=&2\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\Big[p_{jm}-\sigma(\eta_{jm})\Big]\Big[\frac{\partial p_{jm}}{\partial\beta_{\imath\jmath}}+\Delta\sigma'(\eta_{jm})\frac{\partial p_{-jm}}{\partial\beta_{\imath\jmath}}\Big],\\
\frac{\partial\rho}{\partial\theta_k}=&2\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\Big[p_{jm}-\sigma(\eta_{jm})\Big]\Big[\frac{\partial p_{jm}}{\partial\theta_k}-\sigma'(\eta_{jm})(\frac{\partial\xi_{jm}}{\partial\theta_k}-\Delta\frac{\partial p_{-jm}}{\partial\theta_k})\Big],\\
\frac{\partial\rho}{\partial\Delta}=&2\sum_{j=\mathcal{W},\mathcal{K}}\sum_{m=1}^{n}\Big[p_{jm}-\sigma(\eta_{jm})\Big]\Big[\sigma'(\eta_{jm})p_{-jm}\Big],
\end{aligned}$$

where $\eta_{jm} = \xi_{jm} - p^{\beta}_{-j}(\xi_m(\theta))\Delta$,

$$\begin{aligned}
\frac{\partial p_{jm}}{\partial \beta_{\imath\jmath}} =& \sigma'(v_{jm}) s_\imath\big(\sigma(\xi_{jm})\big) s_\jmath\big(\sigma(\xi_{-jm})\big) \\
\frac{\partial p_{jm}}{\partial \theta_k} =& \sigma'(v_{jm})\Big[\sum_{\imath=1}^{K}\sum_{\jmath=1}^{K} \beta_{\imath\jmath}\big(s'_\imath s_\jmath \sigma'(\xi_{jm})\frac{\partial \xi_{jm}}{\partial \theta_k} + s_\imath s'_\jmath \sigma'(\xi_{-jm})\frac{\partial \xi_{-jm}}{\partial \theta_k}\big)\Big].
\end{aligned}$$

Note that $\frac{\partial \ell}{\partial \Delta} = 0$, which implies that the data likelihood is independent of Δ. To make the data likelihood depend on θ more explicitly, we could replace p^{β}_j by $\sigma(\xi_{jm} - p^{\beta}_{-j}(\xi_m(\theta)\Delta))$.

$$\frac{\partial \ell}{\partial \theta_k} = \sum_{j=\mathcal{W},\mathcal{K}} \sum_{m=1}^{n} \left[\frac{d_{jm}}{\sigma_{jm}} - \frac{1-d_{jm}}{1-\sigma_{jm}}\right] \cdot \frac{\partial \sigma_{jm}}{\partial \theta_k}, \tag{A29}$$

$$\frac{\partial \sigma_{jm}}{\partial \theta_k} = \sigma'(\eta_{jm})\Big[\frac{\partial \xi_{jm}}{\partial \theta_k} - \frac{\partial p_{-jm}}{\partial \theta_k}\Delta\Big] \tag{A30}$$

$$\frac{\partial \sigma_{jm}}{\partial \Delta} = \sigma'(\eta_{jm})\Big[- p^{\beta}_{-j}(\xi_m(\theta)\Big] \tag{A31}$$

where $\sigma_{jm} = \sigma(\xi_{jm} - p^{\beta}_{-j}(\xi_m(\theta)\Delta))$ represents the best response to the opponent $-j$.

References

Aguirregabiria, V. and P. Mira (2002): "Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models," *Econometrica*, 70, 1519–1543.

——— (2007): "Sequential estimation of dynamic discrete games," *Econometrica*, 75, 1–53.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010): "Estimating static models of strategic interactions," *Journal of Business & Economic Statistics*, 28, 469–482.

Barwick, P. J. and P. A. Pathak (2015): "The costs of free entry: an empirical study of real estate agents in Greater Boston," *The RAND Journal of Economics*, 46, 103–145.

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

Bresnahan, T. F. and P. C. Reiss (1991): "Empirical models of discrete games," *Journal of Econometrics*, 48, 57–81.

Chen, J., X. Chen, and E. Tamer (2023a): "Efficient estimation in NPIV models: Simulation comparisons of neural network estimators," *Journal of Econometrics*, 235, 1848–1875.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, X., M. L. Gentry, T. Li, and J. Lu (2023b): "Identification and inference in first-price auctions with risk averse bidders and selective entry," *Available at SSRN 3681530*.

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, 2231–2267.

Guerre, E., I. Perrigne, and Q. Vuong (2000): "Optimal nonparametric estimation of first-price auctions," *Econometrica*, 68, 525–574.

Hall, C. A. and W. W. Meyer (1976): "Optimal error bounds for cubic spline interpolation," *Journal of Approximation Theory*, 16, 105–122.

Hall, P., J. Racine, and Q. Li (2004): "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hall, P. G. and J. S. Racine (2015): "Infinite order cross-validated local polynomial regression," *Journal of Econometrics*, 185, 510–525.

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

Iskhakov, F., J. Lee, J. Rust, B. Schjerning, and K. Seo (2016): "Comment on "constrained optimization approaches to estimation of structural models"," *Econometrica*, 84, 365–370.

Jia, P. (2008): "What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry," *Econometrica*, 76, 1263–1316.

Keane, M. P. and K. I. Wolpin (1994): "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence," *The Review of Economics and Statistics*, 648–672.

——— (1997): "The career decisions of young men," *Journal of Political Economy*, 105, 473–522.

Kristensen, D., P. K. Mogensen, J. M. Moon, and B. Schjerning (2021): "Solving dynamic discrete choice models using smoothing and sieve methods," *Journal of Econometrics*, 223, 328–360.

Lee, J. and K. Seo (2015): "A computationally fast estimator for random coefficients logit demand models using aggregate data," *The RAND Journal of Economics*, 46, 86–102.

Li, Q. and J. Racine (2004): "Cross-validated local linear nonparametric regression," *Statistica Sinica*, 485–512.

Luo, Y., I. Perrigne, and Q. Vuong (2018): "Structural analysis of nonlinear pricing," *Journal of Political Economy*, 126, 2523–2568.

Pesendorfer, M. and P. Schmidt-Dengler (2008): "Asymptotic least squares estimators for dynamic games," *The Review of Economic Studies*, 75, 901–928.

——— (2010): "Sequential estimation of dynamic discrete games: A comment," *Econometrica*, 78, 833–842.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, 1403–1430.

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica*, 999–1033.

Shen, X. (1997): "On methods of sieves and penalization," *The Annals of Statistics*, 2555–2591.

——— (1998): "On the method of penalization," *Statistica Sinica*, 337–357.

Stoker, T. M. (1986): "Consistent estimation of scaled coefficients," *Econometrica: Journal of the Econometric Society*, 1461–1481.

Stone, C. J. (1980): "Optimal rates of convergence for nonparametric estimators," *The annals of Statistics*, 1348–1360.

——— (1985): "Additive regression and other nonparametric models," *The annals of Statistics*, 13, 689–705.

Su, C.-L. and K. L. Judd (2012): "Constrained optimization approaches to estimation of structural models," *Econometrica*, 80, 2213–2230.

Sweeting, A. (2013): "Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry," *Econometrica*, 81, 1763–1803.

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

Van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes with Application to Statistics*, New York, Springer.