Market Counterfactuals with Nonparametric Supply: An ML/AI Approach*

Harold D. Chiang† Jack Collison‡ Lorenzo Magnolfi§

Christopher Sullivan¶

November 5, 2025

Abstract

We develop a new approach to market counterfactuals (e.g., merger simulation, tax policy, and product regulation) using machine learning and nonparametric structure from economics. Building on Berry and Haile (2014), we propose a flexible supply specification that relaxes restrictive assumptions about firm conduct and costs. We adapt the Variational Method of Moments (Bennett and Kallus, 2023) with deep neural networks to estimate the model, addressing endogeneity through instrumental variables. Monte Carlo evidence demonstrates good performance even in high-dimensional environments and with moderate sample sizes. Applied to the American Airlines-US Airways merger, our method achieves a fivefold reduction in price prediction error versus standard Bertrand-Nash models.

KEYWORDS: counterfactual analysis, variational method of moments (VMM), neural networks, merger simulation, airline markets

*We thank Dan Ackerberg, Lanier Benkard, Giovanni Compiani, Alessandro Iaria, Phil Haile, Sukjin Han, and the audiences at the Midwest IO Fest 2024, Northwestern, IIOC 2025, Bristol, Warwick, Alpine IO Symposium, and EC '25 for helpful discussions and comments. An extended abstract of a previous version of this work appeared in Proceedings of the 26th ACM Conference on Economics and Computation (EC '25) with the title: "Enhancing the Merger Simulation Toolkit with ML/AI." We thank Dan McLeod for input on previous versions of this project. Any errors are our own.

†Department of Economics, University of Wisconsin-Madison. Email: hdchiang@wisc.edu

‡Department of Economics, University of Wisconsin-Madison. Email: jcollison@wisc.edu

§Department of Economics, University of Wisconsin-Madison. Email: magnolfi@wisc.edu

¶Department of Economics, University of Calgary. Email: christopher.sullivan1@ucalgary.ca

1 Introduction

Economists routinely rely on market counterfactuals to evaluate potential mergers (e.g., Nevo, 2000), taxes (e.g., Miravete, Seim, and Thurk, 2018), product regulation (e.g., Dubois, Griffith, and O'Connell, 2018), and other policy interventions in markets for differentiated products. The standard approach is to perform these counterfactuals using parametric demand and supply models to capture consumers' and firms' responses to policy changes. While recent advances have introduced more flexibility on the demand side (e.g., Compiani, 2022), supply-side modeling remains highly constrained by parametric assumptions about firms' cost and the nature of competition, often defaulting to models of Bertrand-Nash pricing and constant marginal cost. These restrictive assumptions may lead to misleading predictions when the true model of firm conduct and cost differs from the imposed structure. This is not only a theoretical problem: for instance, evidence from merger retrospectives (e.g., Peters, 2006; Björnerstedt and Verboven, 2016) highlights the importance of supply-side assumptions and the potential shortcomings of traditional approaches.

This paper proposes a novel approach that maintains the basic insight that counterfactual outcomes arise in equilibrium after demand and supply responses, while relaxing assumptions about the supply side. Building on the identification framework of Berry and Haile (2014), we introduce a *flexible supply function* that combines markups and marginal costs into a single object that can be estimated without specifying the model of competition. We allow equilibrium prices to depend nonparametrically on market shares, demand derivatives, cost shifters, and ownership structure, thus nesting standard oligopoly models as special cases while letting the data reveal how firms actually set prices. Crucially, we assume demand is known or has been estimated in a first step, making demand derivatives observable and allowing us to focus exclusively on learning supply-side behavior. Identification relies on instrumental variables: intuitively, rival cost shifters provide variation in market shares, while product characteristics excluded from costs help identify how demand derivatives affect pricing. By imposing symmetry across products, we can pool observations within and across markets, making identification feasible. However, even with identification established, the high dimensionality of the problem (with supply depending on market shares and the full matrix of demand derivatives) poses significant estimation challenges.

We overcome these estimation challenges by adapting the Variational Method of Moments (VMM) recently developed by Bennett and Kallus (2023). VMM reformulates the instrumental variables problem as a minimax optimization where the supply function is parameterized as a neural network that grows in complexity with sample size. This approach offers two key advantages for our setting. First, in fixed-dimensional parametric settings, VMM coincides

with the optimally weighted GMM, making it a natural generalization of standard structural methods in industrial organization (IO). Second, deep neural networks can automatically adapt to certain classes of compositional structures in the data (e.g., Bauer and Kohler, 2019; Schmidt-Hieber, 2020), effectively decomposing our high-dimensional problem (e.g., with 30 products, supply depends on over 900 arguments) into lower-dimensional components. Because supply-side behavior may naturally generate such compositional structures, this adaptive capability allows neural network-based VMM to achieve faster convergence rates for our model than traditional nonparametric methods that suffer from the curse of dimensionality.

A key innovation in our adaptation of VMM is to develop procedures to quantify uncertainty in counterfactual predictions. While Bennett and Kallus (2023) provide element-wise inference for individual parameters, market counterfactuals involve complex functions of these parameters, such as equilibrium prices, consumer surplus, or government revenues. We extend the inference framework using the numerical delta method combined with Holm's step-down procedure to construct simultaneous confidence intervals for multiple predictions. This approach not only provides valid uncertainty quantification but also achieves computational efficiency by avoiding the need for test inversions over candidate parameter values. Together, our estimation and inference procedures deliver a practical toolkit: the method completes estimation within hours on standard hardware in typical IO datasets.

We use Monte Carlo simulations to validate the method across several critical dimensions. First, we demonstrate strong predictive accuracy in hold-out samples, with the flexible model achieving mean squared errors close to correctly specified models even with just 100 training markets, while misspecified parametric models generate errors 3-6 times larger. Second, we establish scalability by estimating supply functions in a high-dimensional environment similar to Miller and Weinberg (2017) with 30 products, where the inclusion of demand derivatives and larger network architectures becomes crucial for capturing complex competitive interactions. Third, we verify economic interpretability by showing that the flexible model precisely learns underlying pass-through matrices, including distinctive features like negative cross-price effects under partial internalization that would be missed by misspecified models. These results demonstrate that our approach captures the true economic structure without imposing it, thus addressing a common criticism that views ML/AI approaches as pure statistical "black boxes."

The simulations also validate counterfactual prediction performance in policy-relevant scenarios. For product characteristic regulations that increase the values of one attribute well outside the training support, the flexible model maintains accurate predictions while parametric models show systematic biases. In tax policy analysis, our approach accurately

traces out Laffer curves, correctly identifying revenue-maximizing rates even when extrapolating to tax levels triple those in the training data. For merger simulation, where we consolidate multi-product firms, the flexible model achieves consumer surplus prediction errors less than half those of misspecified models. Finally, our procedure for quantifying uncertainty delivers reliable confidence intervals with 94-98% coverage rates in samples of just 1,000 markets. Together, we take these results as an indication that combining machine learning methods with appropriate economic structure can deliver accurate and interpretable predictions for interesting market counterfactuals.

To demonstrate the practical value of our approach, we apply it to the U.S. airline industry, focusing on the 2013 American Airlines-US Airways merger. The airline industry provides an ideal setting, given its rich variation in market structure across routes and over time, along with evidence of non-Bertrand conduct that makes flexible modeling particularly valuable. Our analysis focuses on markets that transitioned from three to two firms post-merger, where theory suggests the largest competitive effects. The flexible model achieves a 44% improvement in pre-merger fit compared to standard Bertrand-Nash assumptions and, more importantly, delivers much more accurate post-merger predictions. While the Bertrand-Nash model systematically overpredicts price increases, our flexible approach centers its predictions around observed outcomes with a five-fold reduction in passenger-weighted mean squared error (MSE). The improved accuracy is a result of the model's ability to learn actual competitive conduct from the data rather than imposing it, highlighting the practical importance of relaxing standard assumptions in merger evaluation.

While our approach offers significant advantages in flexibility, it comes with important trade-offs that should guide its application. First, without additional restrictions, our baseline method cannot separately identify markups from marginal costs, though Appendix A outlines feasible extensions using market size variation or parametric restrictions that could achieve this decomposition. We leave full development of these extensions to future research, as our primary focus in this paper is on counterfactual prediction. Second, like all nonparametric approaches, the method's predictions for counterfactual scenarios that represent radical departures from observed market structures will be less reliable. However, our simulations demonstrate some robustness in this respect: the flexible model can, in some cases, extrapolate well outside of the support of the data. Finally, the method places greater demands on exogenous identifying variation compared to standard parametric approaches (Magnolfi and Sullivan, 2022). In this sense, our method is complementary to testing approaches that use data to select among a menu of parametric conduct (and cost) models (Backus, Conlon, and Sinkinson, 2021; Duarte, Magnolfi, Sølvsten, and Sullivan, 2024; Dearing, Magnolfi, Quint, Sullivan, and Waldfogel, 2024).

This paper contributes to the IO literature that proposes nonparametric models of market equilibrium. We build our flexible model of supply on the identification results of Berry and Haile (2014), obtaining new results on the nonparametric identification of supply. Similarly to Compiani (2022), who develops a method to estimate demand nonparametrically,[1] our paper proposes a method for nonparametric estimation of the supply side. With similar motivation, recent papers by Gandhi and Houde (2020a) (proposing a linear approximation of the markup function) and Otsu and Pesendorfer (2024) (using the revelation principle) also develop methods to bring more flexible supply models to data. We complement these approaches by proposing a method that leverages advances in ML/AI coupled with a nonparametric structure.

A growing literature seeks to combine ML/AI methods with structure coming from economic theory. Farrell, Liang, and Misra (2020) provide a theoretical foundation for using deep learning in two-stage econometric procedures. Kaji, Manresa, and Pouliot (2023) introduce "adversarial estimation," which formulates the estimator as a minimax problem between a generator and discriminator, achieving parametric efficiency under correct specification. The approach in Wei and Jiang (2025) trains neural networks to directly estimate structural parameters from data moments. Similar to these methods, we aim to leverage the flexibility and computational efficiency of machine learning while preserving the economic structure and interpretability of traditional models. We differ in the specific implementation and application: we formulate a nonparametric structural model of market equilibrium, and estimate it using the VMM formulation of Bennett and Kallus (2023).[2]

The remainder of the paper is organized as follows. Section 2 presents the standard framework that serves as our benchmark. Section 3 develops our flexible approach and discusses identification. Section 4 describes the VMM estimation procedure and derives its statistical properties. Section 5 presents Monte Carlo evidence that compares our method to standard parametric models. Section 6 applies our methodology to airline mergers and discusses the results. Section 7 concludes with a discussion of limitations and directions for future research.

[1] Other recent nonparametric approaches to demand estimation in markets for differentiated products include Tebaldi, Torgovitsky, and Yang (2023) and Brand and Smith (2025).

[2] We also note that VMM is part of a growing recent literature that focuses on applying general function approximation techniques, including deep neural networks and random forests, to instrumental variables problems within a unified framework (see, e.g., Hartford, Lewis, Leyton-Brown, and Taddy, 2017; Dikkala, Lewis, Mackey, and Syrgkanis, 2020; Lewis and Syrgkanis, 2018; Liao, Chen, Yang, Dai, Kolar, and Wang, 2020; Zhang, Imaizumi, Schölkopf, and Muandet, 2023).

2 Market Equilibria and Counterfactuals

This section presents a general equilibrium model for differentiated product markets, following Berry and Haile (2014). We describe the data-generating process, define market counterfactuals, and illustrate with merger simulation as a concrete example.

Data-generating Process: Consumers choose products j from a set $\mathcal{J} = \{1, ..., J\}$ offered by firms across a set of $t \in \mathcal{T} = 1, ..., T$ markets.[3] Each market is characterized by a measure M_t of consumers, which we call market size, and by a $J \times J$ ownership matrix defined as $\mathcal{H}_t = [h_{jkt}]$ with $h_{jkt} = 1$ if the same firm owns products j and k. Each product-market pair (j, t) has a price $p_{jt} \in \mathbb{R}$, a market share $s_{jt} \in (0, 1)$, a vector of product characteristics $x_{jt} \in \mathbb{R}^K$ that enter consumers' demand, and a vector of cost shifters $\mathrm{w}_{jt} \in \mathbb{R}^L$. To streamline notation, for any variable y_{jt}, we denote y_t as the vector of values in market t. Endogenous prices and quantities are generated by equilibrium of demand and supply in market t; we describe these in turn.

Across all markets, the demand system $\mathcal{s}(\cdot) = (\mathcal{s}_1(\cdot), ..., \mathcal{s}_J(\cdot))$ is given by:

$$s_{jt} = \mathcal{s}_j(p_t, x_t, \xi_t), \qquad j = 1, ..., J$$

where $\xi_t = (\xi_{1t}, ..., \xi_{Jt})'$ is a vector of unobservable product characteristics. Define the corresponding matrix of demand derivatives as:

$$D_t \equiv \frac{\partial \mathcal{s}(p_t, x_t, \xi_t)}{\partial p_t'} = \left[\frac{\partial s_{kt}}{\partial p_{jt}}\right]_{j,k=1}^{J}.$$

In general, these derivatives are a function of equilibrium outcomes and exogenous product characteristics, or $D_t = D(p_t, x_t, \xi_t)$.

On the supply side, firm behavior is characterized by a system of first-order conditions for the firms' profit maximization problems:

$$p_{jt} = \Delta_{jt} + c_{jt}, \qquad j = 1, ..., J,$$

where, for each product j, Δ_{jt} is the markup, and c_{jt} is the marginal cost. Markups, which can be expressed as functions $\Delta_{jt} = \Delta_j(s_t, p_t, D(p_t, x_t, \xi_t), \mathcal{H}_t)$, arise endogenously from a model of firm conduct. Firm j's costs are generated by some cost function c_j, or

[3] For expositional simplicity, we keep the set of products offered constant across markets, but none of our results will depend on this assumption.

$c_{jt} = c_j(q_t, \mathrm{w}_{jt}, \omega_{jt})$, where q_{jt} and ω_{jt} are, respectively, equilibrium quantities (obtained as the product of market size and market share, or $q_{jt} = M_t s_{jt}$) and unobserved cost shifter variables. The dependence of costs on q_t allows for the presence of economies of scale and scope. Therefore, a model of conduct and a model of how costs are determined are the key ingredients of the supply-side model.

Example 1. *A canonical assumption in the literature (e.g., in Berry, Levinsohn, and Pakes, 1995) is that the model of conduct is Bertrand-Nash, whereby firms play a complete information pricing game, with constant marginal cost that is a linear index of cost shifters. Under this assumption, market-level markups are* $\Delta_t = \left(\mathcal{H}_t \odot D_t\right)^{-1} s_t$, *where* $\odot$ *denotes the Hadamard product. Furthermore, the cost function is specified as a linear function of observables, or* $c_t = w_t'\gamma + \omega_t$.

To summarize the environment, we have described three functions that pin down the structure of demand ($\mathcal{s}(\cdot)$) and supply ($\Delta(\cdot)$, $c(\cdot)$). The endogenous outcomes (p_t, s_t) are generated by these primitives as a function of exogenous variables $(x_t, \mathrm{w}_t, \xi_t, \omega_t)$. This formulation nests the standard Bertrand-Nash model but allows for more general forms of strategic interaction. With respect to the framework in Section 4.4 in Berry and Haile (2014), the expression above unpacks marginal revenue as the sum of price and markup, which entails no loss of generality.

We restrict the environment with some assumptions on the main primitives. We start with an assumption on equilibrium selection:

Assumption 1. *(Equilibrium Selection)* There exists a unique equilibrium, or the equilibrium selection rule is such that the same p_t arises whenever the vector $(x_t, \mathrm{w}_t, \xi_t, \omega_t)$ is the same.

This assumption, similar to Assumption 13 in Berry and Haile (2014), ensures that prices reflect stable equilibrium behavior. We then impose a mild assumption on cost:

Assumption 2. *(Separability of Cost)* For any product $j = 1, ..., J$, the cost function is separable in unobservable shocks:

$$c_j(q_t, \mathrm{w}_{jt}, \omega_{jt}) = \bar{c}_j(q_t, \mathrm{w}_{jt}) + \omega_{jt}.$$

This separability assumption is essentially without loss of generality because the unobservable component ω_{jt} can be defined as the residual between total costs and the component explained by observables.

We also require the known demand system to satisfy an index restriction on how product characteristics enter demand, which mirrors similar restrictions in Berry and Haile (2014). We partition the product characteristics as $x_{jt} = (x_{jt}^{(1)}, x_{jt}^{(2)})$, where $x_{jt}^{(1)} \in \mathbb{R}$ enters demand only through a linear index with the demand unobservable, while $x_{jt}^{(2)} \in \mathbb{R}^{K-1}$ may enter demand more flexibly.

Assumption 3. *(Index Structure for Demand)* The demand unobservable ξ_{jt} and the characteristic $x_{jt}^{(1)}$ enter preferences only through their sum $\delta_{jt} = x_{jt}^{(1)} + \xi_{jt}$. Therefore, demand, inverse demand, and demand derivatives can be written equivalently as functions of (ξ_t, x_t) or as functions of $(\delta_t, x_t^{(2)})$.

As we show in Section 3.2, this index structure also facilitates identification of supply by creating a lower-dimensional manifold on which the supply function must be identified.

Finally, we restrict the class of supply models we consider, so that prices and product characteristics enter markups only through demand derivatives:

Assumption 4. *(Markup Dependence)* For any product $j = 1, ..., J$, the markup function $\Delta_j(\cdot)$ depends on endogenous market shares s_t and on the matrix of demand derivatives D_t, but conditional on these variables, does not depend on prices p_t or product characteristics (x_t, ξ_t).

Assumption 4 is satisfied by a broad range of conduct models beyond Bertrand-Nash, including Cournot competition, various forms of partial collusion, and models where firms maximize weighted combinations of profits and consumer surplus.[4] Under this assumption we can write markup functions for any $j = 1, ..., J$ as $\Delta_{jt} = \Delta_j(s_t, D_t, \mathcal{H}_t)$.

Observables for the Researcher: In line with standard market data environments, observable variables for the researcher include $(s_t, p_t, x_t, \mathrm{w}_t)$, as well as M_t and $\mathcal{H}_t$. In addition, to identify our model of supply, we will assume that demand is identified, so that the researcher can identify ξ_t and D_t. This echoes standard "two-step procedures" where demand is first estimated or calibrated before estimation or testing of supply.

Assumption 5. *(Known Demand)* The matrix of demand derivatives is known, so that $D_t = D(p_t, x_t, \xi_t)$ is observed.

Market counterfactuals: Most policy changes of interest in applied research can be modeled as *market counterfactuals*. We will predict counterfactuals for a specific set of markets

[4] For a thorough discussion of how the models above (and more) satisfy this assumption, see Appendix C in Dearing et al. (2024).

(periods) $t \in \tilde{\mathcal{T}}$ outside of the sample; here and in what follows, we use tildes to denote counterfactual objects.

A market counterfactual must involve a change in either demand or supply functions, or any of their exogenous arguments. For instance, exogenous product characteristics may now take values $\tilde{x}_{jt}$, or cost functions may be altered to $\tilde{c}_j(\cdot)$. Under the new primitives, consumers and firms react to exogenous changes and counterfactual market outcomes arise from a new market equilibrium $(\tilde{p}_t, \tilde{s}_t)$. These outcomes need to satisfy demand and supply equations, and under our assumptions, can be found as the fixed point of the system:

$$\tilde{p}_{jt} = \tilde{\Delta}_j(\tilde{p}_t, \tilde{D}(\tilde{p}_t, \tilde{x}_t, \tilde{\xi}_t), \tilde{\mathcal{H}}_t) + \tilde{c}_j(\tilde{q}_t, \tilde{\mathrm{w}}_{jt}, \tilde{\omega}_{jt}), \qquad j = 1, ..., J.$$

For any market $t \in \tilde{\mathcal{T}}$, counterfactuals of interests can then be expressed as a map $F(\tilde{p}_t, \tilde{s}_t)$, where we suppress the dependence on (counterfactual) structural objects and exogenous variables. To illustrate the general notion of a market counterfactual, we consider the standard merger simulation problem.

Example 1 (continued)**.** *Suppose the researcher wants to predict prices following a horizontal merger. Further suppose that the true model of conduct is Bertrand-Nash; other assumptions are possible, including Cournot-Nash (e.g., Peters, 2006), Nash Bargaining (e.g., Gowrisankaran, Nevo, and Town, 2015), or profit-weight models (e.g., Miller and Weinberg, 2017). As a baseline case, the merger results only in a deterministic change to the counterfactual ownership matrix $\tilde{\mathcal{H}}_t$ for market t. If the researcher is only interested in predicting counterfactual $(\tilde{p}_t, \tilde{s}_t)$, the map F is the identity. Otherwise, the map F could return other objects of interest, such as counterfactual consumer surplus and firm profit.*

Estimating a Counterfactual: Evaluating the value of F for specific counterfactuals requires predicting endogenous counterfactual outcomes $(\tilde{p}_t, \tilde{s}_t)$, which in turn requires knowledge of the (counterfactual) primitives, exogenous observables, and unobservables. In practice, researchers use a combination of data and assumptions. Typically, the functions $\mathfrak{s}$ and $\bar{c}$ are estimated with parametric models, and markup functions $\Delta_j(\cdot)$ follow from an assumption on conduct. This allows the researcher to obtain estimates of in-sample ξ_t and ω_t, which are used to specify counterfactual $\tilde{\xi}_t$ and $\tilde{\omega}_t$. Counterfactual $\tilde{\mathfrak{s}}(\cdot), \tilde{c}(\cdot)$, and $\tilde{\Delta}(\cdot)$ are either kept fixed, or changed deterministically.

Example 1 (continued) **.** *Standard merger simulation (e.g., Werden and Froeb, 1994; Nevo, 2000) typically proceeds in three steps. First, the researcher formulates a parametric (e.g.,*

linear, logit, or mixed logit) demand system $\mathcal{s}(\cdot;\theta^D)$ *and estimates demand primitives* $\hat{\theta}^D$*, which imply values of and* $\hat{\xi}_t$ *and demand derivatives* $\hat{D}_t = D(p_t, x_t, \xi_t; \hat{\theta}^D)$ *in each market* t*. Second, the researcher assumes a supply model, which encompasses models of cost and markups. In the standard toolkit, these are typically a constant marginal cost function, and Bertrand-Nash conduct. Marginal costs can thus be inverted as* $\hat{c}_t = p_t - \left(\mathcal{H}_t \odot \hat{D}_t\right)^{-1} s_t$*. Third, the researcher computes post-merger prices* $\tilde{p}_t$ *in each market* t *under the post-merger ownership structure* $\tilde{\mathcal{H}}_t$ *holding everything else fixed. Post-merger prices are thus the fixed point of the nonlinear system:*

$$\tilde{p}_t = \hat{c}_t + \left(\tilde{\mathcal{H}}_t \odot D(\tilde{p}_t, x_t, \hat{\xi}_t; \hat{\theta}^D)\right)^{-1} \mathcal{s}(\tilde{p}_t, x_t, \hat{\xi}_t; \hat{\theta}^D). \tag{1}$$

The same procedure applies to alternative assumptions on cost and conduct. For instance, assumptions on cost efficiencies generated by the merger, quantified in post-merger cost vectors $\tilde{c}_t$*, can be incorporated by using this vector instead of* $\hat{c}_t$ *in Equation* (1)*.*

When performing market counterfactuals in applications, researchers trade off practicality and data limitations (which suggest tight parametric specifications) with the dangers of misspecification (which may lead to misleading conclusions). Compiani (2022), building on the nonparametric demand framework in Berry and Haile (2014), shows how to nonparametrically estimate demand, and what the consequences of restrictions on demand are for counterfactuals. Complementary to these efforts, we propose an approach to flexibly estimating supply to relax the parametric restrictions on costs and markups.

3 Flexible Supply for Market Counterfactuals

3.1 The Flexible Supply Function

Building on the equilibrium framework in Section 2, we develop a flexible approach to modeling supply that maintains economic structure while relaxing parametric assumptions. Under Assumptions 1-4, equilibrium prices satisfy the system of equations $p_{jt} = \Delta_j(s_t, D_t, \mathcal{H}_t) + \bar{c}_j(M_t s_t, \mathrm{w}_{jt}) + \omega_{jt}$ for all $j = 1, \ldots, J$, where markup functions $\Delta_j(\cdot)$ depend on market shares and demand derivatives, and marginal cost functions $\bar{c}_j(\cdot)$ depend on quantities $q_{jt} = M_t s_{jt}$ and cost shifters. These relationships lead to the following definition.

Definition 1 (*Flexible Supply Function*)**.** For each product $j \in \{1, \ldots, J\}$, the *flexible supply*

function h_j is the sum of the markup and marginal cost functions:

$$h_j(s_t, D_t, \mathrm{w}_{jt}, M_t, \mathscr{H}_t) \equiv \Delta_j(s_t, D_t, \mathscr{H}_t) + \bar{c}_j(M_t s_t, \mathrm{w}_{jt}), \qquad j = 1, ..., J.$$

The flexible supply function allows us to express the equilibrium price equation as a nonparametric regression model with an additive, unobservable shock:

$$p_{jt} = h_j(s_t, D_t, \mathrm{w}_{jt}, \mathscr{H}_t, M_t) + \omega_{jt}, \qquad j = 1, ..., J. \tag{2}$$

This equation forms the basis of our estimation strategy. The primary restriction in this formulation is that it does not impose separability between markups and costs. Although these functions can be recovered nonparametrically with additional restrictions, we do not pursue this approach in our main specification. Appendix A provides details on this extension.

Our main specification, especially when coupled with a symmetry restriction, allows us to efficiently exploit both across- and within-market variation present in standard IO data, thus lending itself to practical approaches to identification and estimation. Moreover, our flexible supply function accommodates several important classes of counterfactuals common in applied economics, where researchers typically estimate a demand system and then simulate market outcomes under an assumed model of supply.

We will formally consider identification of the flexible supply functions in Equation (2) before discussing the symmetry restriction and feasible counterfactuals.

3.2 Identification

Identification of the functions h_j in Equation (2) requires addressing the endogeneity of its arguments, (s_t, D_t), which are simultaneously determined with prices. In turn, identification requires valid instrumental variables. For notational simplicity in the following arguments, we condition on a fixed market size M_t and ownership structure $\mathscr{H}_t$, suppressing them as arguments of the function h_j. The identification results hold conditional on any given values of these market-level variables.

Following the nonparametric identification literature (Newey and Powell, 2003; Berry and Haile, 2014), the basis for identification is the conditional moment restriction:

$$\mathbb{E}[\omega_{jt} \mid z_{jt}, \mathrm{w}_{jt}] = \mathbb{E}[p_{jt} - h_j(s_t, D_t, \mathrm{w}_{jt}) \mid z_{jt}, \mathrm{w}_{jt}] = 0, \tag{3}$$

where z_{jt} denotes instruments for product j in market t, and w_{jt} represents own exogenous cost shifters that enter the supply function directly.

Two special challenges arise in this setting. First, we must ensure we identify supply rather than inverse demand. For example, under logit demand where $D_t = D(s_t)$, the supply function $h_j(s_t, \mathrm{w}_{jt})$ could potentially replicate inverse demand $p_{jt} = [\log s_{jt} - \log s_{0t} - x_{jt}^{(1)} - \xi_{jt}]/\alpha$. We thus impose exclusion restrictions to prevent this confusion:

Assumption 6 (*Exclusion Restrictions*)**.** There exists a vector of instruments z_{jt} satisfying $\mathbb{E}[\omega_{jt} \mid z_{jt}, \mathrm{w}_{jt}] = 0$, where z_{jt} contains the full vector of product characteristics x_{jt}, and $x_{jt}^{(1)}$ is excluded from the cost shifters w_{jt}.

Second, without additional structure, we would need to identify product-specific functions h_j of $J + J^2$ endogenous variables; finding instruments that independently move all the variables (s_t, D_t) as required by standard completeness conditions (Newey and Powell, 2003) would be extremely demanding, if not infeasible in most applications.[5] Our key insight is that s_t and D_t are linked through the (known) demand function. Thus, each function h_j needs only to be identified on the manifold:

$$\mathcal{M} = \{(s_t, D_t) : D_t = D(s_t, \delta_t, x_t^{(2)}) \text{ for some } (\delta_t, x_t^{(2)}) \in \mathrm{Supp}(\delta_t, x_t^{(2)} | s_t)\},$$

where $\delta_t = x_t^{(1)} + \xi_t$ from Assumption 3. This manifold has dimension at most $2J + J(K-1)$, which is (in typical applications) lower than the $J + J^2$ dimensions of the full space. We therefore require a completeness condition adapted to the manifold structure:

Assumption 7 (*Manifold Completeness*)**.** For any measurable function $B : \mathcal{M} \times \mathcal{W} \to \mathbb{R}$ with finite expectation:

$$\mathbb{E}[B(s_t, D_t, \mathrm{w}_{jt}) \mid z_{jt}, \mathrm{w}_{jt}] = 0 \text{ a.s.} \implies B(s_t, D_t, \mathrm{w}_{jt}) = 0 \text{ a.s. on } \mathcal{M}.$$

Under these assumptions, we can prove that the functions h_j are identified:

Theorem 2 (*Identification of Supply*)**.** *Under Assumptions 1-7, the supply functions h_j are identified on $\mathcal{M}$ for all $j = 1, \ldots, J$.*

Proof. See Appendix B. □

As is typical with completeness conditions, the exact requirements on the economic primitives that generate completeness are not transparent. We therefore provide a discussion in Appendix B that includes more constructive identification arguments and a comparison with

[5] Even exploiting the fact that D_t is symmetric for many commonly used demand systems, we would still need $O(J^2)$ instruments in that case.

the inverse supply identification results in Berry and Haile (2014).

Instrumental Variables: The identification strategy requires instruments that provide two types of variation. With endogenous variables (s_t, D_t) constrained to lie on manifold $\mathcal{M}$, we need excluded instruments only for (s_t, δ_t). Rival cost shifters $\mathrm{w}_{-j,t}$ serve as natural instruments for market shares: when competitor costs change, their equilibrium prices adjust, which in turn affects own market share through substitution patterns. Meanwhile, excluded product characteristics $x_{jt}^{(1)}$ provide variation in demand and its derivatives without directly affecting costs.

The identification strategy can also be illustrated via a practical approach using predicted instruments, similar in spirit to the predicted prices procedures in Berry, Levinsohn, and Pakes (1999) and Gandhi and Houde (2020b), though with a fundamental difference. While those papers use predicted prices to identify demand given a known supply model, here we use predicted demand derivatives to identify supply given known demand. Specifically, we can construct predicted market shares $\hat{s}_t$ and predicted demand derivatives $\hat{D}_t$ using the projection of endogenous variables on instruments and the known functional form of demand. These predicted values lie on the manifold $\mathcal{M}$ by construction and can serve as generated instruments in estimation.

Crucially, for each product-specific function h_j, we observe only one realization per market—the equilibrium outcome for product j in market t. Thus, identification of h_j relies entirely on variation in $(s_t, D_t, \mathrm{w}_{jt})$ across markets. With J products and T markets, we effectively have T observations to identify each of the J functions. We address this challenge with symmetry restrictions.

3.3 Exploiting Within-Market Variation Through Symmetry

As anticipated in Section 3.1, we now introduce a symmetry restriction to develop a powerful and practical estimation framework. While Theorem 2 establishes nonparametric identification, estimating J separate, high-dimensional functions can be challenging in finite samples. To improve statistical power by pooling information across products, we restrict all firms to share the same underlying conduct model and cost function, though they may differ in their demand conditions and cost realizations.

Assumption 8. *(Supply Symmetry)* The supply function satisfies an exchangeability restriction such that for all products $j = 1, \ldots, J$:

$$h_j(s_t, D_t, \mathrm{w}_{jt}, \mathcal{H}_t, M_t) = h(s_{jt}, s_{-j,t}, D_{jt}, D_{-j,t}, \mathrm{w}_{jt}, \mathcal{H}_t, M_t),$$

where $s_{-j,t}$ denotes the vector of rival market shares and D_{jt} and $D_{-j,t}$ partition the demand derivative matrix into own and cross-price derivatives.

Under symmetry, data from all $J \times T$ product-market observations can be pooled to estimate a single function h, rather than J separate functions. This allows us to leverage within-market variation to learn supply.

The symmetry assumption is satisfied by a broad class of standard oligopoly models. Under Bertrand-Nash competition, all firms maximize profits, taking rivals' prices as given, yielding identical markup functions up to the reordering of arguments. Similarly, Cournot competition generates symmetric quantity-setting behavior across firms. Models of symmetric collusion, where firms place identical weights κ on rivals' profits as in profit-weight models, also satisfy this restriction, as shown in the following example.

Example 3. *Consider the profit-weight model with common conduct parameter κ and constant marginal cost. Each firm f maximizes $\pi_f + \kappa \sum_{g \neq f} \pi_g$, where π_g denotes profits for firm g, yielding first-order conditions that, when reordered according to ownership, take the same functional form across all products. Specifically, the markup for product j owned by firm f is:*

$$\Delta_j = e_j'(\mathcal{H}_f \odot D_t)^{-1} s_{f,t}$$

where $\mathcal{H}_f$ has elements equal to 1 for products owned by f and κ for rival products, $s_{f,t}$ contains shares ordered by ownership, and e_j is the appropriate selection vector. After reordering, this yields the same functional form for all products.

However, the symmetry assumption rules out important classes of models with heterogeneous firm conduct or cost. Examples include leader-follower dynamics as in Stackelberg competition, which involve fundamentally asymmetric strategic considerations that cannot be captured under Assumption 8.[6]

3.4 Discussion: Scope and Limitations

Our framework accommodates a wide range of counterfactuals, including three broad classes of particular interest in the applied literature:

1. *Cost and Product Regulations*: Environmental standards, safety requirements, and quality mandates through changes in cost shifters w_{jt} or product characteristics x_{jt}.

[6] The practical implementation of Assumption 8 for markets with varying numbers of firms and products requires a data pre-processing step to reorder and pad the input vectors to ensure a consistent input dimension for estimation—see Appendix C for a constructive example.

Examples include emissions standards (Goldberg, 1998) and carbon pricing policies where regulatory costs enter as observable shifters.

2. *Tax and Subsidy Policies*: Unit taxes, ad valorem taxes, and subsidies through adjustments to the supply relation. The framework can be used to compute equilibrium pass-through, tax incidence, optimal rates, and government revenue (Miravete et al., 2018; Berry et al., 1999).

3. *Market Structure Changes*: Mergers, acquisitions, divestitures, and product entry/exit through modifications to ownership matrices $\mathcal{H}_t$ and product sets $\mathcal{J}$. Applications include horizontal merger evaluation (Miller and Weinberg, 2017; Gowrisankaran et al., 2015) and new product welfare analysis (Petrin, 2002).

However, there are important caveats to keep in mind when using the flexible supply model. First, without additional restrictions (see Appendix A), we cannot separately identify the levels of markups and costs, which in turn prevents measurement of firm profits and markups. This also prevents counterfactuals where the cost or markup functions are altered in a known way, e.g., if a merger leads to coordinated effects or a change in the degree of scale economies.

Second, while Section 3.2 provides a good sense of the variation that is needed to identify the model, our approach based on a flexible supply function will naturally require richer exogenous identifying variation compared to standard parametric approaches. In this sense, we believe our method is complementary to testing approaches that select among parametric models (e.g., Backus et al., 2021; Duarte et al., 2024), which are a less data-intensive alternative way of introducing flexibility in the supply side (Magnolfi and Sullivan, 2022).

Finally, as with any nonparametric approach, there will be limitations in predictions of counterfactual scenarios that represent radical departures from observed market structures. We address this latter point with simulations in Section 5.

4 Estimation

We now turn to the estimation of the flexible markup function using neural Variational Method of Moments (Bennett and Kallus, 2023), which we refer to as VMM throughout. We first formally define the VMM estimator, then discuss the rationale for adopting VMM, and finally outline the procedures for pointwise and simultaneous inference based on the VMM estimates.

To introduce VMM, explicitly, we use a moment condition for structural markup:

$$\mathbb{E}[p_{jt} - h_j(s_t, D_t, \mathrm{w}_{jt}; \theta, \mathscr{H}_t) \mid z_t, \mathrm{w}_{jt}] = 0$$

Denote $N = TJ$. Given a preliminary consistent estimate $\tilde{\theta}_N$, the estimator solves a min-max program:

$$\hat{\theta}_N = \mathrm{argmin}_{\theta\in\Theta}\mathrm{sup}_{f\in\mathscr{F}_N} \frac{1}{TJ}\sum_{j,t} f(z_{jt})'\omega_{jt}(\theta) - \frac{1}{4TJ}\sum_{j,t}(f(z_{jt})'\omega_{jt}(\tilde{\theta}_N))^2 - R_N(f,h) \quad (4)$$
$$\text{where} \quad \omega_{jt}(\theta) = p_{jt} - h_j(s_t, D_t, \mathrm{w}_{jt}; \theta, \mathscr{H}_t)$$

Following the identification arguments in Section 3, the estimator includes a vector of observable cost shifters w_{jt} and a vector of instruments z_{jt}.[7] The instruments are necessary to address the endogeneity of market shares s_t and demand derivatives D_t. In general, $f \in \mathscr{F}_N$ and $h \in \mathscr{H}_N$ are adequately chosen sequences of function classes. For neural VMM, we use classes of neural networks with growing width and depth, allowing flexible controls of model complexity to cope with the curse of dimensionality. R_N is a regularizer $R_N : \mathscr{F}_N \times \mathscr{H}_N \to [0, \infty]$ that penalizes the complexity of the neural network; we note in Appendix A.3 that regularization can also be used to adhere to economic properties of the markup function. θ is a potentially large set of parameters that pins down the function h. Under regularity conditions, Theorem 4 of Bennett and Kallus (2023) establishes the consistency of the estimator in Equation (4) for the true parameter values θ_0.

4.1 Discussion of VMM

We now elaborate on our choice of the VMM estimator. This choice is motivated by two key considerations: first, its encompassment of the conventional Generalized Method of Moments (GMM); and second, the adaptive capability of deep neural network architectures to capture underlying structural features and alleviate the curse of dimensionality.

In the parametric setting, VMM bears a close relationship to GMM—the standard workhorse of structural estimation in economics. When the parameter space is fixed and finite-dimensional, VMM coincides with the optimally weighted GMM estimator (OWGMM) in the absence of additional regularization. This equivalence underscores VMM's role as a natural generalization of traditional parametric estimators, extending their applicability to high-dimensional and nonparametric environments while maintaining interpretability within

[7] In practice, we exclude observable cost shifters from the set of instruments. Given that the function f is highly nonlinear, the cost shifters no longer explicitly instrument for themselves. Simulations reveal better estimates upon the exclusion of w from the set of instruments.

conventional parametric frameworks.

To illustrate the connection between traditional OWGMM estimators and VMM, we reproduce the proof of Lemma 1 from Bennett and Kallus (2023), which succinctly illustrates the intuition behind the terms in the VMM objective function in Equation (4). Let $\tilde{\theta}_N$ denote a preliminary estimate of the parameters (e.g., from the first step of OWGMM), and let Γ_N be the optimal weighting matrix based on $\tilde{\theta}_N$. We claim that

$$\hat{\theta}^{\text{OWGMM}}(f_1, \ldots, f_k, \tilde{\theta}_N) = \hat{\theta}_N^{\text{VMM}}(\text{span}\{f_1, \ldots, f_k\}, \tilde{\theta}_N).$$

By definition of OWGMM,

$$\begin{aligned}
\hat{\theta}_N^{\text{OWGMM}}(f_1, \ldots, f_k, \tilde{\theta}_N) &= \arg\min_{\theta \in \Theta} \left\| \Gamma_N^{-1/2} \mathbb{E}_N[f(z)\omega(x;\theta)] \right\|^2 \\
&= \arg\min_{\theta \in \Theta} \sup_{v \in \mathbb{R}^k} \left\{ v' \mathbb{E}_N[f(z)\omega(x;\theta)] - \tfrac{1}{4} v' \Gamma_N v \right\} \\
&= \arg\min_{\theta \in \Theta} \sup_{v \in \mathbb{R}^k} \left\{ \mathbb{E}_N[(f(z)'v)'\omega(x;\theta)] - \tfrac{1}{4} \mathbb{E}_N[((f(z)'v)'\omega(x;\tilde{\theta}_N))^2] \right\}.
\end{aligned}$$

The first equality follows from the dual norm representation, and the second uses the fact that $f(z)'v \in \text{span}\{f_1, \ldots, f_k\}$.

On the other hand, although conventional Nonparametric Instrumental Variable (NPIV) methods are well understood (see the reviews of, e.g., Chen 2007; Carrasco, Florens, and Renault 2007) and provide substantial flexibility in modeling complex relationships, these methods are severely constrained by the curse of dimensionality, which limits their practical applicability to low-dimensional settings. In contrast, deep neural network-based estimators, such as the proposed neural VMM, have the potential to circumvent this curse by exploiting adaptable architectures that align with underlying structural regularities. As a nonparametric method, VMM inherits the advantages of conventional NPIV—most notably, its ability to accommodate complex dependence between endogenous and exogenous variables—while alleviating dimensionality challenges.

Deep neural networks differ from traditional series-based methods in that their architecture allows them to automatically adapt to lower-dimensional structures whenever such structures are present (e.g., Bauer and Kohler 2019; Schmidt-Hieber 2020). In such cases, the network effectively decomposes a high-dimensional estimation problem into a sequence of lower-dimensional subproblems, thereby mitigating the curse of dimensionality. Consequently, deep neural network-based estimators can potentially achieve substantially faster approximation when the target function exhibits a compositional structure that consists of functions with lower dimensional arguments, a common situation in structural economic ap-

plications. A detailed discussion is provided in Appendix D, following the setup in Schmidt-Hieber (2020).

4.2 Quantification of Uncertainty

We now turn to the quantification of uncertainty in the flexible markup function. For the asymptotic analysis, we consider regimes in which J remains fixed while $N \to \infty$ (equivalently, $T \to \infty$).

While Bennett and Kallus (2023) offers valid element-wise testing procedures for the underlying parameters, our focus is on predicting counterfactual prices — a complex functional of these parameters. To address this, we develop an inference procedure that employs the numerical delta method in conjunction with Holm's step-down procedure and a permutation procedure, as well as proposing a novel algorithm to construct simultaneous confidence intervals. This approach enables the quantification of uncertainty in counterfactual price predictions through standard errors and confidence intervals while ensuring computational feasibility.

We now describe the estimation procedure. Formal statements of the theoretical results are provided in Appendix E. Suppose that a consistent first-stage estimator $\tilde{\theta}_N$ is available. For any sufficiently smooth transformation h, standard regularity conditions imply that:

$$\left\{\nabla_{\theta'} h(\theta_0, \mathcal{H}) \Omega_0^{-1} \nabla_{\theta'} h(\theta_0, \mathcal{H})'/N\right\}^{-1/2} \left(h(\hat{\theta}_N, \mathcal{H}) - h(\theta_0, \mathcal{H})\right) \xrightarrow{d} N(0, I).$$

This asymptotic result is infeasible for direct application, as the asymptotic variance in this expression is generally complex and unknown to the researcher. Note that $\nabla_\theta h(\theta_0; \mathcal{H})$ has dimension $d \times b$. In the simplest case, suppose that $d = 1$ and the parameter of interest is $h_x(\theta_0, \mathcal{H})$. Lemma 9 in Bennett and Kallus (2023) states that, for any $\beta \in \mathbb{R}^b$, we have:

$$\beta^T \Omega_0^{-1} \beta = -\frac{1}{4} \inf_{\gamma \in \mathbb{R}^b} \sup_{f \in \mathcal{F}} \mathbb{E}[f(z)' \nabla_\theta \omega(\theta_0) \gamma] - \frac{1}{4} \mathbb{E}[(f(z)' \omega(\theta_0))^2] - 4\gamma' \beta - R_N(f). \tag{5}$$

Taking $\beta = \nabla_\theta h_x(\theta_0, \mathcal{H})$ for x, the solution to the optimization problem above yields the asymptotic variance in Equation (E1). The gradient $\nabla_\theta h_x(\theta_0, \mathcal{H})$ is difficult to compute analytically but numerical differentiation[8] can be employed, e.g., Hong et al. (2015), and $\hat{\theta}_N$ can be used in place of θ_0.[9]

[8] For numerical differentiation to be valid, the ϵ in the difference has to satisfy $\epsilon \to 0$ and $N\epsilon / \log N \to \infty$ from Theorem 1 in Hong, Mahajan, and Nekipelov (2015).

[9] In practice, we use automatic differentiation in `torch`. Broadly speaking, the module stores a computational graph as the neural network is fit; training requires and stores the gradients in backpropagation. We note that we are still subject to the regularity condition from Hong et al. (2015).

However, this approach cannot directly obtain a covariance matrix when $d \geq 2$. We extend the method to provide a simultaneous confidence interval by adapting Holm's Step-Down procedure (Holm, 1979) with the estimates for $\hat{\sigma}^2_{x_j}(\hat{\theta}, \mathcal{H})$ and $h_{x_j}(\hat{\theta}, \mathcal{H})$ for each $j = 1, ..., d$. The set of critical values T_α is known for significance levels $\frac{\alpha}{d+1-k}$ with $k = 1, ..., d$. For any ordering of x and fixed ordering T_α, we can compute the (vectorized) confidence interval $h_x(\hat{\theta}, \mathcal{H}) \pm N^{-\frac{1}{2}} \hat{\sigma}_x(\hat{\theta}, \mathcal{H}) T_\alpha$. To implement the procedure, we construct the intervals for all permutations of $j = 1, ..., d$, resulting in $d!$ permutations of x. This is because we must consider any possible ordering of the p-values of $x_1, ..., x_d$. The simultaneous confidence interval is subsequently the union of the bounds in each of the $d!$ permutations. Formally, the proposed procedure is as follows:

Algorithm 1 Simultaneous Confidence Interval for neural VMM

1: **for** each index $j \in \{1, ..., d\} \equiv J$ **do**
2: Estimate $\hat{\sigma}^2_{x_j}(\hat{\theta}, \mathcal{H})$ for $\sigma^2_{x_j}(\theta_0, \mathcal{H})$ at x_j by solving Equation 5
3: **end for**
4: Fix critical values $T_\alpha = \left\{ T_{\alpha_k} : k = 1, ..., d \right\}$ where $\alpha_k = \frac{\alpha}{d+1-k}$
5: **for** each permutation $\tilde{J}$ of indices J **do**
6: Arrange values $\tilde{x}$ and $\hat{\sigma}_{\tilde{x}}$ using permuted indices $\tilde{J}$
7: Construct bounds as $h_{\tilde{x}}(\hat{\theta}, \mathcal{H}) \pm N^{-\frac{1}{2}} \hat{\sigma}_{\tilde{x}}(\hat{\theta}, \mathcal{H}) T_\alpha$ for fixed T_α
8: **end for**
9: Collect the simultaneous confidence interval as the union of bounds from Steps 5-8

This procedure allows us to quantify uncertainty simultaneously across many predicted prices. Importantly, we can construct confidence intervals on counterfactuals using our method. This algorithm not only provides simultaneous inference via VMM by integrating the numerical delta method with Holm's step-down and an efficient permutation-based implementation, but it also sidesteps the need for the econometrician to conduct infinitely many test inversions over candidate parameter values. This feature substantially enhances computational efficiency, making our method both robust and practical for high-dimensional counterfactual analysis.

In applied economic analysis, prices are seldom the only counterfactual objects of interest. We therefore extend the procedure to encompass smooth functionals of prices. Examples 4-6 below illustrate the construction of simultaneous confidence intervals for market-level counterfactual quantities and for functionals aggregated across markets, such as total output, total consumer surplus, and total government revenue. Formal derivations, additional examples, and details are provided in Appendix E.

Example 4 (Counterfactual market share for a single product)**.** *Consider inference on the counterfactual market share of product j, defined by the functional $F(h(\theta)) = s_j(h(\theta))$. We have:*

$$\left\{D_j(h(\theta_0))\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)' D_j(h(\theta_0))'/N\right\}^{-1/2}\left(F(h(\hat{\theta}_N)) - F(h(\theta_0))\right) \xrightarrow{d} N(0, I),$$

where $D_j(h(\theta))$ denotes the j-th row of the Jacobian of the demand system, evaluated at the counterfactual price vector $h(\theta)$.

Example 5 (Total consumer surplus across markets)**.** *Consider the total consumer surplus across markets $m = 1, ..., M$ from logit demand with the map:*

$$F(h(\theta)) = -\frac{1}{\alpha}\sum_m \log\left(1 + \sum_{j\in\mathcal{J}_m}\exp(-\alpha h_j(\theta) + x_j\beta + \xi_j)\right).$$

We have:

$$\left\{s(h(\theta_0))'\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)' s(h(\theta_0)))/N\right\}^{-1/2}\left(F(h(\hat{\theta}_N)) - F(h(\theta_0))\right) \xrightarrow{d} N(0, I),$$

where $s(h(\theta))$ are market shares evaluated at counterfactual prices $h(\theta)$.

Example 6 (Government revenue from ad valorem taxes)**.** *Denote $\mathcal{J} = \mathcal{J}_1 \cup \cdots \cup \mathcal{J}_M$ as the union of product sets across markets $m = 1, ..., M$. Consider government revenue from ad valorem taxes ν across markets with the map $F(h(\theta)) = \sum_{j\in\mathcal{J}} \nu_j h_j(\theta) s_j(h(\theta))$. We have:*

$$\left\{G(h(\theta_0))\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)' G(h(\theta_0))'/N\right\}^{-1/2}\left(F(h(\hat{\theta}_N)) - F(h(\theta_0))\right) \xrightarrow{d} N(0, I),$$

where $G(h(\theta)) = (\nu \odot s(h(\theta))' + (\nu \odot h(\theta))' D(h(\theta))$ and $D(\cdot)$ is the matrix of stacked block-diagonal by markets.

5 Monte Carlo Simulations

We conduct extensive Monte Carlo simulations that go beyond standard finite-sample assessments to evaluate whether our flexible approach can serve as a practical tool for policy analysis. Our simulations address five critical dimensions of performance. First, we examine predictive accuracy in hold-out samples across a range of realistic market sizes, from small datasets ($T = 100$) to large-scale applications ($T = 10,000$). Second, we test scalability by implementing the method in high-dimensional environments with 30 products and over 900

supply function arguments. Third, we demonstrate economic interpretability by showing that the model correctly learns pass-through relationships. Fourth, we evaluate counterfactual predictions for product regulations, tax policies, and merger simulations applications where the model must extrapolate beyond training support. Finally, we verify that our inference procedure delivers reliable confidence intervals, providing useful uncertainty quantification.

5.1 Simulation Environments

In order to fully illustrate the dimensions of performance discussed above, we develop three environments that vary in complexity. In each environment, we generate data under four supply-side specifications (crossing two conduct models with two cost structures) at multiple sample sizes. For conduct, we consider both the Bertrand-Nash pricing model and a profit-weight model in which firms place a weight κ on rivals' profits. In terms of cost, we consider specifications where marginal cost is either constant or decreasing in quantity. This 2×2 crossing yields four distinct data-generating processes per environment.[10]

Baseline Environment. Markets feature 2-3 single-product firms. On the demand side, we adopt a logit specification. When considering profit-weight models, we set $\kappa = 0.5$. We generate samples with $T \in \{100, 1000, 10000\}$ markets.

High-Dimensional Environment. To address concerns about the curse of dimensionality, we implement a second environment inspired by Miller and Weinberg (2017) featuring 30 differentiated products per market — an order of magnitude larger than our baseline. This yields a $30 \times 30 = 900$ dimensional demand derivative matrix. We adopt a more flexible nested logit demand system than in our baseline environment and generate samples with $T \in \{100, 1000\}$. When considering profit weight models, we set $\kappa = 0.75$.

Merger Simulation Environment. For merger counterfactuals (Section 5.5), we augment the high-dimensional environment with richer ownership variation. Markets contain either three firms (with 6, 5, and 4 products) or four firms (one each with 5 and 4 products, two with 3 products), with 0-5 products randomly dropped per market to yield 10-15 active products. This ensures that the pre-merger data contains variation in ownership ($\mathcal{H}_t$) analogous to the merger we simulate, helping the flexible model learn how the proposed change in market structure will affect equilibrium pricing. When considering profit-weight models, we set $\kappa = 0.75$. We use $T = 1,000$ markets.

[10] Appendix F fully describes these environments and all simulation details.

Crucially for identification, markets in all data environments feature variation in both observed cost shifters w_{jt} and product characteristics x_{jt} excluded from cost, allowing us to form the necessary instrumental variables as discussed in Section 3.2. Our simulation strategy, summarized in Appendix Table F1, lists the specific data environments we use when evaluating each of the five dimensions of our model's performance.

5.2 Hold-out Sample Performance

We begin by examining the ability of our flexible model to predict prices in hold-out samples, i.e., markets not used during estimation. This exercise provides a direct test of whether the learned supply function captures market outcomes beyond the training data in our baseline environment with 2-3 products per market. We randomly sample 80% of markets for training and reserve the remaining 20% as a hold-out test sample.[11] For each hold-out market, we predict prices using the estimated flexible supply function and compare these to predictions from standard parametric models.

The implementation details for both the flexible and parametric models, including choice and formation of instruments and computation of counterfactuals, are provided in Appendix F. For the flexible model, the primary implementation decision involves choosing the neural network architecture. We experiment with three different network dimensions: a "small" 3×3 hidden layer, a "medium" 20×20 hidden layer, and a "large" 100×100 hidden layer specification. We also report results from a flexible model that omits demand derivatives D_t as an argument to assess their importance. Appendix G reports additional results.

Table 1 reports mean squared errors (MSE)[12] for price predictions when data are generated under two different assumptions on conduct in the DGP, corresponding to panels A and B respectively, and a cost function with economies of scale.[13] This setting is particularly challenging as it requires the flexible model to learn how both markups and marginal costs vary with quantities without imposing parametric structure.[14]

For each sample size $T \in \{100, 1000, 10000\}$, we first report the MSE for the correctly specified parametric model (e.g., Bertrand-Nash with economies of scale under the corresponding DGP in Panel A), which achieves the lowest MSE and represents the irreducible er-

[11] Within the training sample, 20% is reserved as a hold-out validation set used for early stopping.

[12] The mean squared error for price predictions is computed as $\text{MSE} = \frac{1}{N_h} \sum_{t \in \mathcal{T}_h} \sum_{j=1}^{J_t} (\hat{p}_{jt} - p_{jt})^2$, where $\hat{p}_{jt}$ denotes the predicted price and p_{jt} the true price for product j in market t, $\mathcal{T}_h$ is the hold-out test sample, J_t is the number of products in market t, and $N_h = \sum_{t \in \mathcal{T}_h} J_t$ is the total number of product-market observations in the hold-out sample.

[13] Results for constant marginal cost DGPs are similar and reported in Appendix Table G1.

[14] While in this table we report MSE results from a single simulation, in Table G2 we report quantiles of the distribution of MSEs across 100 simulations where models are retrained with different weight initializations. The ranges are relatively tight, suggesting results are not sensitive to simulation error and initialization.

TABLE 1: MSE Across Models, Baseline Environment with Economies of Scale

T	True Model	Standard Models			Flexible Models			D_t included
		B	M	P	$h = 3$	$h = 20$	$h = 100$	
Panel A: Bertrand-Nash DGP								
100	0.95	2.05	478.57	5.41	1.32	1.20	0.96	No
					1.09	1.11	0.99	Yes
1,000	0.98	2.63	603.44	9.45	1.30	1.31	1.24	No
					1.39	1.39	1.29	Yes
10,000	0.99	2.93	926.42	9.54	1.43	1.44	1.08	No
					1.39	1.42	1.07	Yes
Panel B: Profit-Weight DGP								
100	0.95	3.78	33.85	6.82	1.48	1.55	1.08	No
					1.30	1.39	1.32	Yes
1,000	0.98	5.85	32.45	9.45	1.94	1.97	1.15	No
					1.57	1.41	1.05	Yes
10,000	0.99	6.40	46.43	9.73	2.12	1.71	1.07	No
					1.66	1.05	1.03	Yes

Notes: The table reports the mean squared error (MSE) in prices for the true model, a set of standard parametric models (B = Bertrand with constant cost, M = joint profit maximization with constant cost, P = marginal cost pricing with constant cost), and the flexible model. In Panel A, the true supply model generating the data is Bertrand with economies of scale, while in Panel B, it is a profit-weight model with economies of scale. Each DGP is from the single-product baseline environment. The Flexible Model columns include neural networks of varying dimensions (e.g., $h = 3$ refers to a 3×3 hidden layer). The MSE in predicted prices is computed on a hold-out test sample.

ror from cost shocks. This provides a natural benchmark for evaluating our flexible approach.

Several patterns emerge from these results. First, the flexible model achieves MSEs close to the irreducible error level while significantly outperforming misspecified parametric models. For instance, in Panel B where data are generated under profit-weight conduct with economies of scale, the flexible model with $T = 10,000$ and a large network achieves an MSE of 1.03 (nearly matching the true model's 0.99) while the misspecified Bertrand model has an MSE of 6.40, over six times larger. The misspecified monopoly and perfect competition models perform even worse, with MSEs of 46.43 and 9.73 respectively.

Second, the results demonstrate good finite-sample performance. Even with just $T = 100$ markets, the flexible model outperforms misspecified parametric alternatives, though sample size improvements are moderate in this relatively simple environment—MSEs decrease gradually from around 1.3-1.5 at $T = 100$ to 1.0-1.1 at $T = 10,000$ for the best-performing flexible specifications.

Third, in this baseline environment with logit demand, including demand derivatives D_t provides only modest improvements, consistent with the relatively simple substitution

patterns.[15] Similarly, while larger networks ($h = 100$) perform somewhat better than smaller ones, even the small 3×3 network substantially outperforms misspecified parametric models. This suggests that in simpler market settings, model flexibility matters more than network capacity or derivative information.

Remark 1 (Robustness to Misspecification). The flexible model achieves comparable performance to correctly specified parametric models and substantially outperforms misspecified models across all data-generating processes, even with moderate sample sizes (e.g., $T = 1,000$).

5.3 Scalability to High-Dimensional Environments

A fundamental limitation of nonparametric methods is the curse of dimensionality. In our application, the input dimensionality of the supply function grows quadratically with the number of products J because it includes both market shares $s_t \in \mathbb{R}^J$ and demand derivatives $D_t \in \mathbb{R}^{J \times J}$. To test whether our flexible model estimated with VMM can overcome this challenge, we perform simulations in an environment mimicking Miller and Weinberg (2017). With a richer demand system and $J = 30$ products (an order of magnitude larger than in our baseline setting), there are over 900 demand derivative arguments. These derivatives take a more complex form than those in our baseline environment.

Table 2 presents results from this high-dimensional environment. In this complex setting, the importance of model capacity and derivative information becomes dramatically apparent. Under the Bertrand DGP (Panel A), the flexible model with $T = 1,000$ and demand derivatives achieves an MSE of 1.28, close to the true model's 1.07, while substantially outperforming misspecified models (monopoly: 6.28, perfect competition: 3.55).

The value of demand derivatives is particularly striking under the profit-weight DGP (Panel B), where strategic complementarities interact with complex substitution patterns. With $T = 1,000$, including derivatives reduces MSE from 5.17 to 1.67 for the small network and from 4.08 to 1.71 for the medium network—improvements of over 60%. Larger networks also become crucial: with derivatives, the large network achieves an MSE of 1.57 compared to 1.67-1.71 for smaller networks. These results highlight that as market complexity increases through richer demand systems, more products, and strategic interactions, both network capacity and derivative information become essential for capturing equilibrium relationships.

These results demonstrate that VMM successfully handles high-dimensional problems: the network learns which of the more than 900 potential interactions matter for pricing,

[15] Under logit demand, demand derivatives can be constructed with a simple function of only pairs of market shares. Intuitively, the demand derivative matrix becomes redundant because the neural network can learn D_t from just s_t.

TABLE 2: MSE Across Models, High-dimensional Environment with Constant Costs

T	True Model	Standard Models			Flexible Models			D_t included
		B	M	P	$h = 3$	$h = 20$	$h = 100$	
Panel A: Bertrand-Nash DGP								
100	1.15	1.15	5.07	2.86	1.40	1.40	1.41	No
					1.41	1.42	1.44	Yes
1,000	1.07	1.07	6.28	3.55	1.36	1.16	1.12	No
					1.31	1.28	1.28	Yes
Panel B: Profit-Weight DGP								
100	1.15	2.74	3.35	2.89	3.01	3.17	2.72	No
					2.53	2.67	2.64	Yes
1,000	1.07	5.08	5.64	5.67	5.17	4.08	1.60	No
					1.67	1.71	1.57	Yes

Notes: The table reports the mean squared error (MSE) in prices for the true model, a set of standard parametric models (B = Bertrand with constant cost, M = joint profit maximization with constant cost, P = marginal cost pricing with constant cost), and the flexible model. In Panel A, the true supply model generating the data is Bertrand with constant marginal cost, while in Panel B, it is a profit-weight model with constant marginal cost. The Flexible Model columns include neural networks of varying dimensions. The MSE in predicted prices is computed on a hold-out test sample.

guided by the moment conditions. This scalability opens new possibilities for counterfactual analysis in retail markets with rich product variety, where traditional nonparametric methods would face challenges.

Remark 2 (Overcoming Dimensionality). The flexible model estimated with VMM maintains good predictive performance with $J = 30$ products, achieving MSEs within $20 - 50\%$ of the correctly specified model, while misspecified parametric models have MSEs $3-5$ times larger.

5.4 Interpretation via Pass-Through Analysis

A common criticism of ML/AI methods in economics is their "black box" nature: while they may predict well, the economic mechanisms underlying these predictions remain opaque. To address this concern, we examine whether our flexible model learns economically meaningful supply-side relationships by analyzing the implied pass-through matrices.

Pass-through matrices capture how cost shocks propagate through equilibrium prices, capturing both direct effects of own-cost changes and strategic responses to rivals' price adjustments. These matrices provide insight into the economic structure learned by our model. We compute pass-through by increasing costs by 1 percent, solving for new equilibrium prices, and calculating the corresponding percentage price changes.

Table 3 presents pass-through matrices for a representative duopoly market under dif-

TABLE 3: Simulated Pass-through Matrices

Panel A: Bertrand-Nash DGP				Panel B: Profit-Weight DGP			
True Model		Flex. Model		True Model ($\kappa = 0.5$)		Flex. Model	
0.41	0.03	0.48	0.10	0.38	−0.28	0.40	−0.18
0.13	0.91	0.09	0.73	0.03	0.95	0.00	0.94

Notes: The table reports the simulated pass-through matrices of the true models and the flexible supply model. For a representative duopoly market t, element P_{ij} of the matrix represents a percentage change in p_{it} after a 1% change in cost loaded on ω_{it}. Panel A represents matrices obtained under a Bertrand-Nash DGP for the true model and for the estimated flexible supply model trained on $T = 1,000$ markets. Panel B represents matrices under a profit-weight DGP ($\kappa = 0.5$).

ferent data-generating processes in our baseline environment.[16] We train the flexible model using $T = 1,000$ markets, using a medium size network ($h = 20$), and including as an argument the matrix D_t. Panel A shows results under Bertrand-Nash competition with constant marginal cost, while Panel B presents results under profit-weight conduct. The flexible model's implied pass-through matrices closely match the truth in both cases.

Under Bertrand-Nash competition (Panel A), the flexible model estimates own-cost pass-through of 0.48 and 0.73 on the diagonal, compared to 0.41 and 0.91 for the true model. Cross-cost pass-through values are 0.10 and 0.09 for the flexible model versus 0.03 and 0.13 for the true model. While not exact, these estimates correctly capture the key economic features: substantial own-cost pass-through and positive but smaller cross-cost effects reflecting strategic complementarities.

Under profit-weight conduct (Panel B), our estimated flexible model implies own-cost pass-through of 0.40 and 0.94 (versus true values of 0.38 and 0.95). Notably, the flexible model correctly identifies negative cross-cost pass-through (−0.18 versus true value of −0.28), a distinctive feature of partial collusion where firms internalize effects on rivals' profits. This sign reversal from the Bertrand-Nash case demonstrates that the flexible model can learn different modes of conduct without any direct assumptions on the game-theoretic structure.

Remark 3 (Economic Interpretability). The flexible model learns the basic economics of supply-side relationships as summarized by pass-through matrices, correctly identifying positive cross-effects of cost increases under Bertrand-Nash competition and negative cross-effects under profit-weight conduct.

[16] We show results here for the median duopoly market by inside share. The results are robust to the choice of market.

5.5 Market Counterfactuals

We now turn to simulations that showcase the core application of our method: predicting market outcomes under counterfactual scenarios. Following the framework developed in Section 2, we examine three classes of counterfactuals that span the range of policy interventions commonly studied in empirical IO: (i) product regulation through modification of characteristics, (ii) tax policy design via Laffer curve analysis, and (iii) merger simulation through changes in ownership structure.[17] For each counterfactual, we solve for new equilibrium prices under both the flexible model and various parametric specifications, obtain counterfactual objects under different transformations F, and then compare predictions to the true counterfactuals. Throughout, we focus on the following key issue: how far out of the support of the data can our flexible, data-driven model give useful counterfactual predictions?

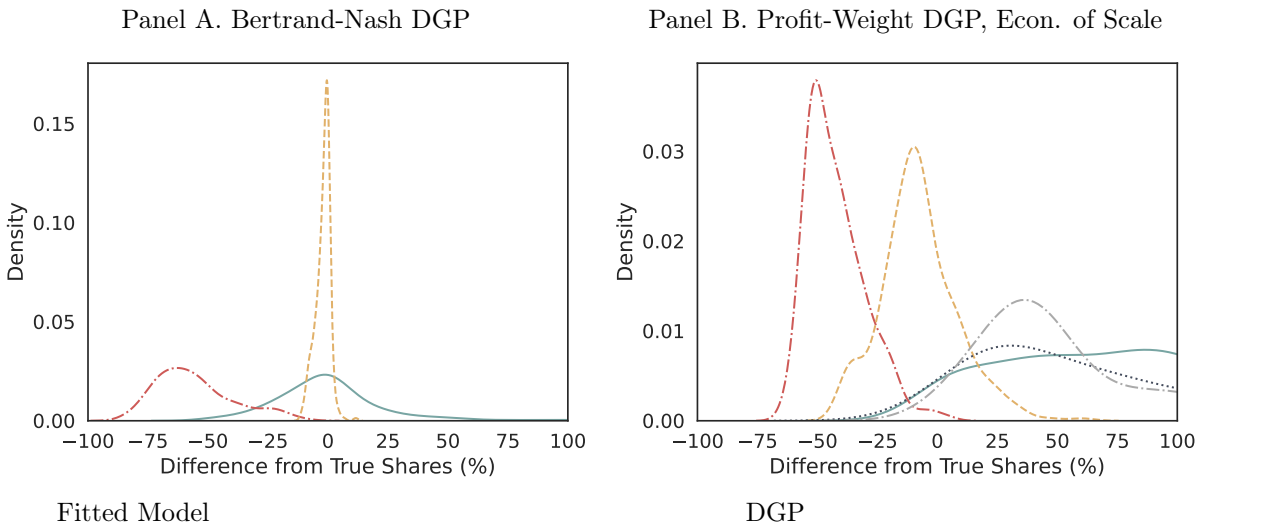
Product Characteristic Regulation: Governments frequently regulate product characteristics to influence consumer behavior, from fuel economy and air pollution standards (Ito and Sallee, 2018; Jacobsen, Sallee, Shapiro, and Van Benthem, 2023) to sugar content limits. We examine counterfactuals where, in our baseline environment, characteristic x_1 is shifted up by one. This pushes the counterfactual well beyond the training support where $x_1 \sim U[0,1]$, thus testing the flexible model's ability to predict equilibrium responses to product changes not observed in the data.

Figure 1 shows that the flexible model accurately predicts consumption (market share) changes due to regulations on product characteristics, achieving small absolute error across DGPs and beating misspecified parametric models. For instance, under the profit-weight DGP with economies of scale (Panel B), the flexible model achieves a root mean percentage squared error (RMPSE) of 17.57%,[18] while all parametric models achieve RMPSEs above 40%. Moreover, the flexible model's distribution of prediction errors is tightly centered around zero, while misspecified models show systematic biases.

[17] Appendix G reports additional results, including robustness exercises, other counterfactuals, and inference. The other counterfactuals we consider are regulations on product characteristics that affect cost and product exit.

[18] This is defined as $\text{RMPSE} = 100 \times \sqrt{\frac{1}{N_c} \sum_{t,j} \left(\frac{\tilde{s}^m_{jt} - \tilde{s}_{jt}}{\tilde{s}_{jt}} \right)^2}$, where $\tilde{s}^m_{jt}$ denotes the predicted counterfactual share under model m, $\tilde{s}_{jt}$ is the true counterfactual share, and N_c is the total number of product-market observations in the counterfactual.

FIGURE 1: Regulation of Product Characteristics: Share Predictions



	A. Bertrand	B. Profit-Weight	C. Bertrand (Scale)	D. Profit-Weight (Scale)
Bertrand (Scale)	-	-	-	77.3
Bertrand (Const.)	-	53.51	14.27	100.58
Monopoly	57.88	51.22	58.26	44.08
Perf. Comp.	23.29	59.72	24.94	77.59
Flex Supply	3.65	10.07	11.59	17.57

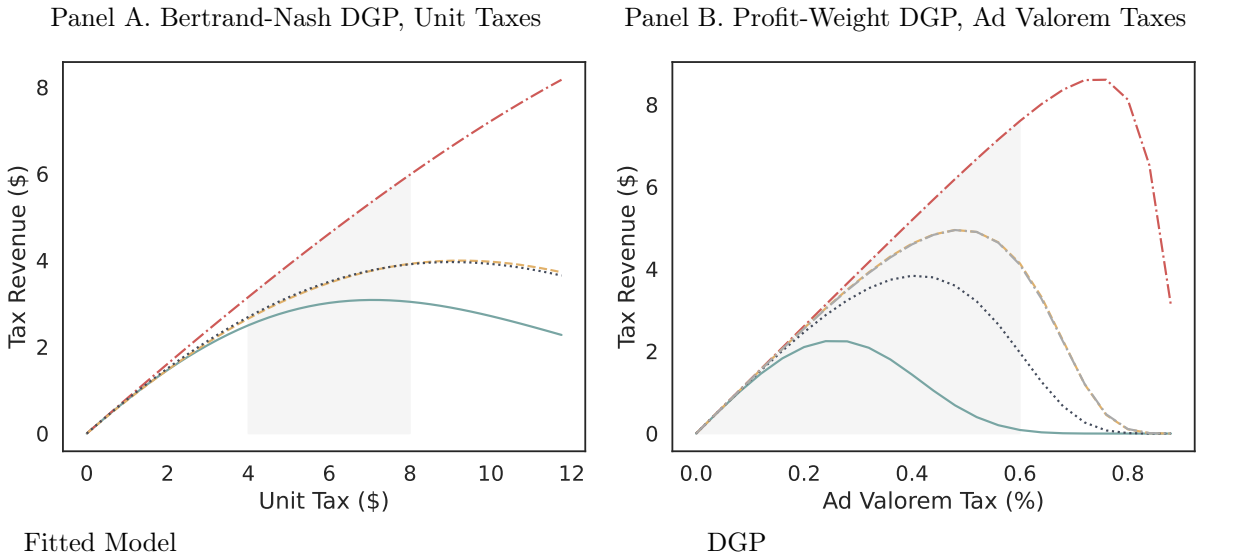
Notes: The figure displays share prediction errors when characteristic $\tilde{x}_1 = x_1 + 1$ is pushed out of its training support. The flexible model uses a medium neural network with demand derivatives on $T = 1,000$ markets.

Tax Policy (Laffer Curves): Tax policy design requires understanding the Laffer curve, or how revenues vary with tax rates (e.g., Miravete et al., 2018). This is a highly nonlinear relationship that depends critically on firm conduct and cost structure. We examine both unit taxes τ (levied on firms) and ad valorem taxes ν (levied on consumers), as formalized in Equation (F1) in Appendix F. Training data include tax variation with $\nu_t \sim U[0, 0.6]$ and $\tau_{jt} \sim U[4, 8]$, but we evaluate predictions at rates up to 90% for ad valorem and \$12 for unit taxes, well outside the training support.[19]

Figure 2 demonstrates the flexible model's ability to capture the nonlinear relationship between tax rates and revenues. In Panel A (Bertrand DGP, unit taxes), the flexible model's Laffer curve virtually overlays the true curve, correctly identifying the revenue-maximizing rate around \$9. The misspecified perfect competition model substantially underestimates revenues at all rates, predicting a revenue-maximizing rate around \$6. For ad valorem taxes

[19] In Appendix Figure G6, we reduce the variation in tax rates found in the training data. The results show that variation in ad valorem taxes is especially useful for learning the Laffer curve.

FIGURE 2: Laffer Curves for Unit and Ad Valorem Taxes



Fitted Model	DGP			
	A. Bertrand (U)	B. Profit-Weight (U)	C. Bertrand (AV)	D. Profit-Weight (AV)
Bertrand	-	0.51	-	1.02
Monopoly	2.07	0.99	3.28	3.82
Perf. Comp.	0.75	1.19	2.12	2.41
Flex Supply	0.04	0.04	0.15	0.02

Notes: The figure displays predicted Laffer curves under different DGPs and different estimated models. The flexible model uses a medium neural network ($h = 20$) with demand derivatives, trained on $T = 1,000$ markets in the baseline environment augmented with taxes. Shaded areas correspond to the support of tax rates in training data.

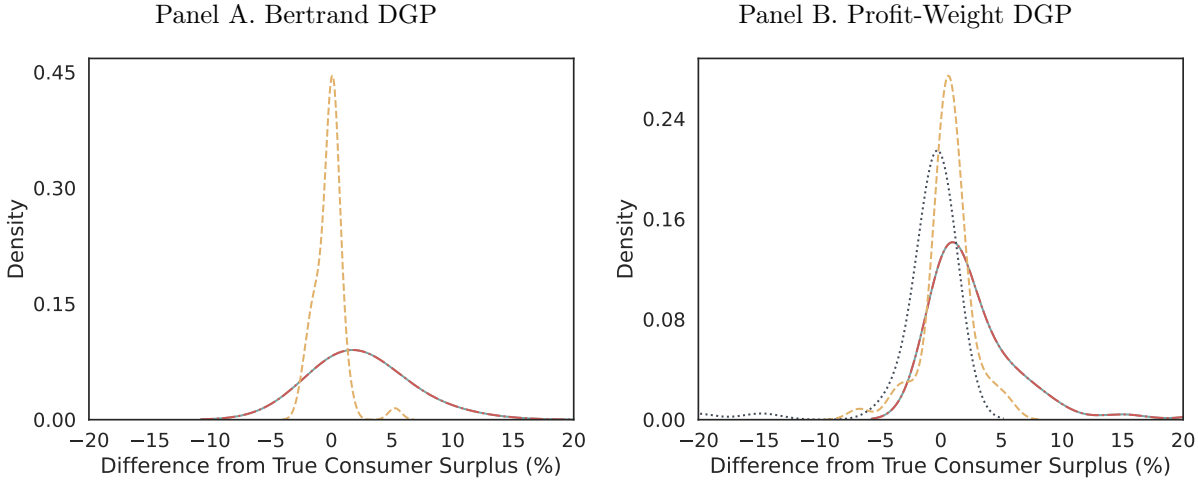
under profit-weight conduct (Panel B), the flexible model achieves an MSE of just 0.02 in revenue predictions,[20] compared to 1.02 for Bertrand-Nash and 3.82 for monopoly.

Merger Simulation: Merger evaluation represents a prominent counterfactual in applied work, and thus an important application of our method. Following the standard approach outlined in Example 1 of Section 2, we model mergers as changes in the ownership matrix $\mathcal{H}_t$ while maintaining all products in the market. Our merger simulation environment features two market structures: half contain three multi-product firms (with 6, 5, and 4 products, respectively), while half contain four firms (one each with 5 and 4 products, and two with 3 products each). We introduce additional variation by randomly dropping 0-5 products per

[20]For revenue predictions, MSE is computed as MSE $= \frac{1}{K}\sum_{k=1}^{K}(\hat{R}_k - R_k)^2$, where $\hat{R}_k$ and R_k are predicted and true revenues at tax rate k, and K is the number of tax rates evaluated. We use MSE rather than RMPSE for Laffer curves because revenue approaches zero at high counterfactual taxes, rendering percentage differences unstable due to divide-by-zero issues.

market, yielding 10-15 active products.[21] The merger counterfactual consolidates the two three-product firms in four-firm markets, reducing competition from four to three firms and creating a six-product merged entity. This setup makes our counterfactual relatively close to the support of the data, allowing the flexible model to learn how ownership concentration affects pricing across diverse market configurations.

FIGURE 3: Merger Simulation: Consumer Surplus Changes



Fitted Model	Panel DGP: A. Bertrand	Panel DGP: B. Profit-Weight
Bertrand (Const.)	-	3.71
Monopoly	8.69	5.02
Perf. Comp.	8.69	5.02
Flex Supply	1.17	2.08

Notes: The figure displays the distribution of percentage differences between predicted and true consumer surplus changes from multi-product mergers. Each curve represents the kernel density of prediction errors (predicted minus true, divided by true, in percentage terms) across test markets. The table reports root mean squared errors (RMSE) of consumer surplus predictions. The flexible model uses a medium-sized neural network (20×20 hidden layer) with demand derivatives as inputs, trained on $T = 1,000$ markets. Additional results for product exit counterfactuals are presented in Appendix G.2.

Figure 3 presents the distribution of prediction errors for consumer surplus changes from mergers. The panels display the percentage deviation of each model's predicted consumer surplus change from the true change, with the accompanying table reporting root mean percentage squared errors (RMPSE) for these predictions. Under the Bertrand DGP (Panel A),

[21] This induces a sparse composite structure in the DGPs, under which sparse neural network-based VMM can potentially achieve faster approximation rates than conventional nonparametric estimators; see Appendix D for details.

the flexible model achieves an RMPSE of 1.17, substantially outperforming the misspecified monopoly and perfect competition models (both with RMPSE of 8.69). The performance advantage is similarly pronounced under the profit-weight DGP (Panel B), where the flexible model's RMPSE of 2.08 is less than half that of the misspecified parametric models (3.71 for Bertrand, 5.02 for both monopoly and perfect competition).

The distribution plots reveal that the flexible model's predictions are tightly centered around zero error, while misspecified models show systematic biases. Under Bertrand competition (Panel A), the monopoly and perfect competition models substantially overestimate consumer harm, while under profit-weight conduct (Panel B), the misspecified Bertrand model shows a wider distribution of errors.

Other Counterfactuals: Beyond the three primary counterfactuals examined above, our framework readily extends to additional policy-relevant scenarios. In Appendix G.2, we present results for product exit counterfactuals, where products are removed from the market due to regulatory bans or firm decisions. The flexible model accurately predicts the resulting price increases and consumer welfare losses, capturing both the direct effects of reduced competition and the reallocation of demand across remaining products. We also examine counterfactuals involving changes in marginal cost shifters (beyond tax policy), such as input price shocks or technological improvements. As shown in Appendix Table G2, the flexible model maintains prediction accuracy even for cost changes of 50-100% beyond the training support, demonstrating its ability to capture the underlying cost pass-through structure. These additional exercises confirm that our approach provides a general-purpose tool for counterfactual analysis, capable of addressing a wide spectrum of "what-if" questions that arise in applied economics.

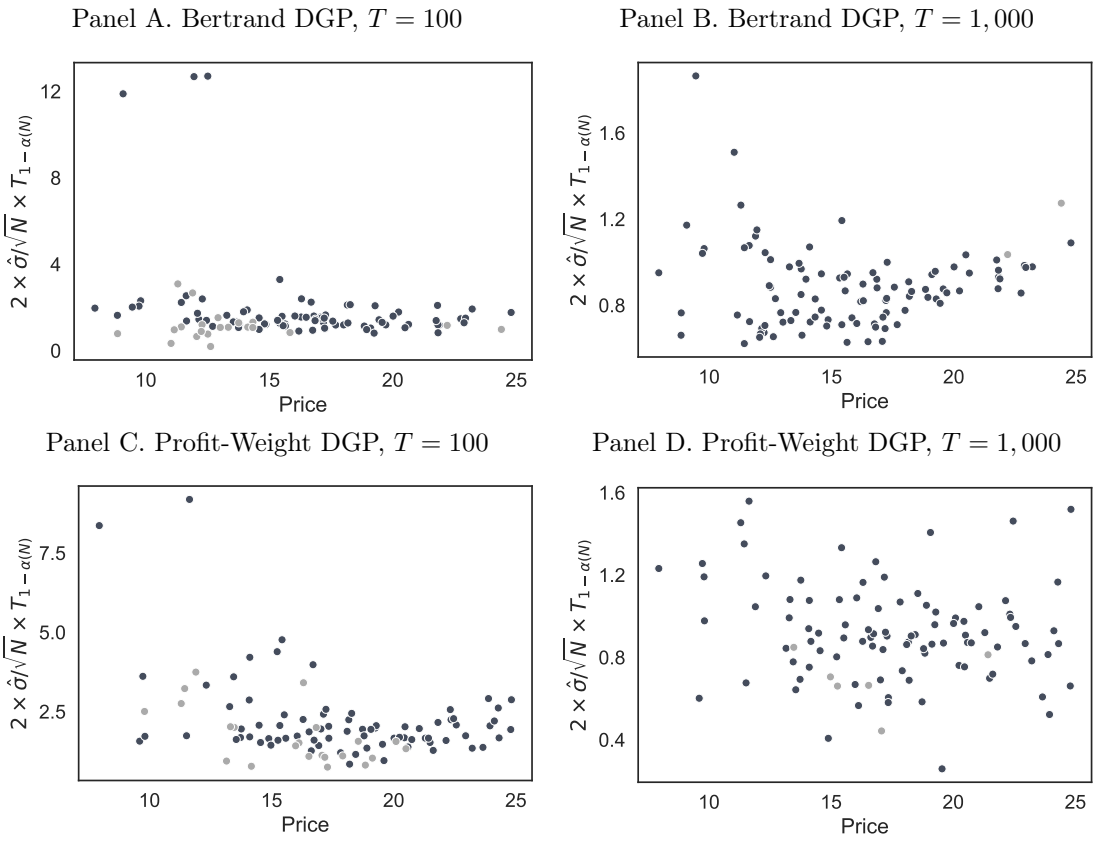
Remark 4 (Extrapolation Capability). The flexible model performs well in counterfactuals, maintaining prediction accuracy substantially beyond the training support, while misspecified parametric models show systematic biases regardless of distance from support.

5.6 Quantification of Uncertainty in Prediction

Beyond point predictions, quantifying uncertainty is essential for policy decisions. We evaluate the finite-sample performance of our inference procedure using a product exit counterfactual, where one product is removed and we predict the resulting equilibrium prices with 95% confidence intervals. Figure 4 displays confidence interval widths (computed using the delta method with Bonferroni correction) against predicted prices across 100 test markets, examining both small ($T = 100$) and typical ($T = 1,000$) sample sizes under Bertrand and

profit-weight DGPs.

FIGURE 4: Inference on Counterfactual Merger Simulation Prices



	$T = 100$		$T = 1,000$	
	A. Bertrand	B. Profit-Weight	C. Bertrand	D. Profit-Weight
Within	79%	78%	98%	94%
Outside	21%	22%	2%	6%

Notes: The figure displays 95% confidence interval widths for counterfactual prices following a product exit. Each point represents the confidence interval width (computed as $2 \times \hat{\sigma}/\sqrt{N} \times T_{1-\alpha}(N)$ with Bonferroni correction) for a single product's predicted price across 100 test markets. The table reports coverage rates, i.e., the percentage of true counterfactual prices falling within the predicted confidence intervals, decomposed into "within" (prices inside the training support) and "outside" (prices beyond the training support). The flexible model uses a medium-sized neural network (20×20 hidden layer) with demand derivatives. Panels A and B show results under Bertrand conduct with constant costs for $T = 100$ and $T = 1,000$ training markets, respectively. Panels C and D show results under profit-weight conduct ($\kappa = 0.5$) with economies of scale for the same sample sizes. Appendix Figure G11 presents results without demand derivatives.

The results demonstrate strong inference properties that improve markedly with sample size. With $T = 1,000$ markets, our procedure achieves 94-98% coverage rates, closely matching the nominal 95% level, while confidence interval widths shrink substantially compared to $T = 100$. Importantly, the method maintains reliable coverage both within the training support (exceeding 94% for $T = 1,000$) and when extrapolating beyond it (up to 94% for

profit-weight conduct), though extrapolation coverage is more variable. The heteroscedastic pattern—wider intervals at price extremes—reflects data density in the training sample and appropriately captures greater uncertainty where fewer observations are available. These results confirm that VMM provides not only accurate counterfactual predictions but also allows for informative uncertainty quantification crucial for policy evaluation.

6 Empirical Application

As a showcase of our method, we examine the US airline industry. The airline industry has received substantial attention from the IO literature (starting with Berry, 1992; Berry, Carnall, and Spiller, 2006) given the rich available data and significant consolidation over the last two decades. The retrospective studies of large mergers have had mixed results (Peters, 2006), potentially linked to non-Bertrand conduct (see, e.g., Ciliberto and Williams, 2014, for evidence of non-competitive conduct). Our method applies well here given the large amount of data and variation in market structure.[22]

6.1 Background and Data

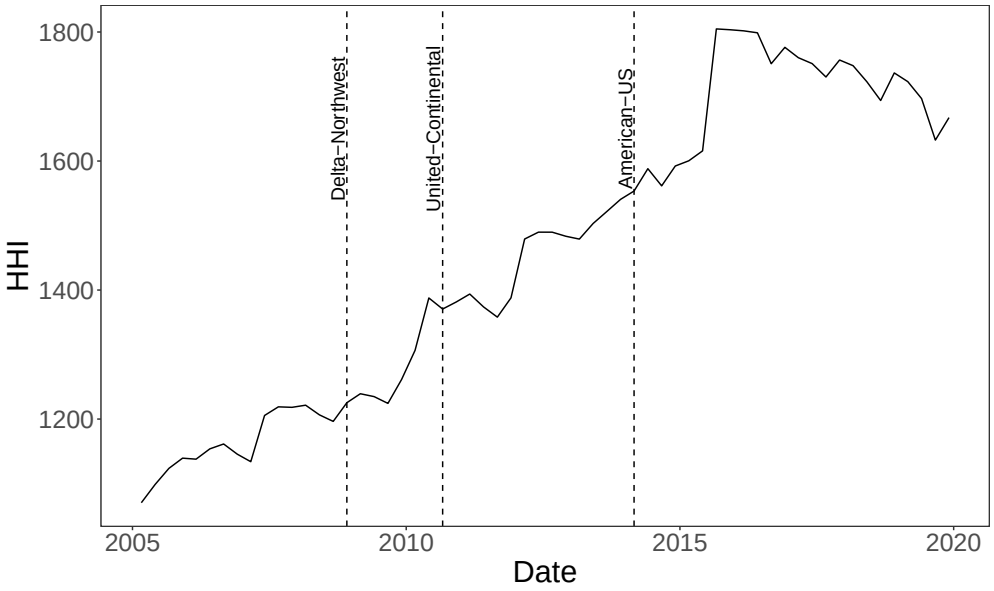
We construct a database of the US airline industry during the period 2005-2019. We use the 10 percent sample of purchased airline tickets from the well-known Airline Origin and Destination Survey (DB1B) database released by the US Department of Transportation. Following Azar, Schmalz, and Tecu (2018) and Kennedy, O'Brien, Song, and Waehrer (2017), a market is defined as a pair of cities, regardless of the flight direction. We match cities to Metropolitan Statistical Areas (MSA) and collect data on the populations of these MSAs from the Bureau of Economic Analysis. The geometric mean of endpoint populations is used as a measure of the market size. A product is a one-way trip that services a particular city-pair and is defined at the carrier-market-quarter level. Additional details on the construction of the data can be found in Appendix H.1.

The U.S. airline industry has experienced substantial consolidation in the last two decades with legacy carriers and low-cost airlines participating in large mergers. We show a descriptive increase in the Herfindahl-Hirschman Index (HHI) in Figure 5. The earliest merger in our data is the Delta-Northwest merger in 2008. The merger was announced on April 14,

[22] We emphasize the illustrative nature of our application. Recent papers have highlighted the dynamic nature of pricing and demand in this market (e.g., Williams, 2022; Betancourt, Hortaçsu, Oery, and Williams, 2022; Aryal, Murry, and Williams, 2024; Hortaçsu, Natan, Parsley, Schwieg, and Williams, 2024), and the role of endogenous network structure (Ciliberto, Murry, and Tamer, 2021; Li, Mazur, Park, Roberts, Sweeting, and Zhang, 2022; Bontemps, Gualdani, and Remmy, 2023; Yuan and Barwick, 2024). We abstract away from these important elements to keep our application tied to the standard merger simulation toolkit.

2008, and was approved on October 29, 2008, after roughly six months of review by the US Department of Justice (DOJ). Given the limited overlap between the merging airlines' operations, the merger was perceived as having a modest impact on competition. The second merger included is the United-Continental merger in 2010. The DOJ approved the merger after four months of review on August 27, 2010. As a condition of approval, the merged entity was required to lease slots to Southwest at Newark Liberty Airport in New Jersey. Finally, we consider the controversial merger of American Airlines and US Airways. The last of the "mega-mergers" that involved two airlines, the deal was announced on November 12, 2013. According to the settlement terms, the merged entity was required to divest slots at several major airports, most prominently at Ronald Reagan Washington National Airport and New York's LaGuardia Airport. More recently, outside our analysis, Alaska Airlines acquired Virgin America and Hawaiian Airlines, and a federal court blocked JetBlue's attempted acquisition of Spirit.

FIGURE 5: Concentration in the Airline Industry



Notes: The figure plots the evolution of the national HHI of the airline industry during the period 2005-2019. We use passenger counts to construct market shares used in the calculation of HHI.

6.2 Demand and Supply Estimation

Demand Estimation: We follow Berry and Jia (2010) in adopting a nested logit demand model. We briefly summarize the model here and provide additional details in Appendix H.2. Product characteristics include average fares, the share of nonstop flights, the average

distance in thousands of miles, and a squared distance term. We restrict our attention to the major carriers, controlling for the number of fringe firms to capture variation in market structure over time across origin-destination pairs. We include origin-destination fixed effects. Our nesting structure includes all inside goods in one nest. We include instruments to handle endogeneity issues for prices and nests. We use BLP instruments as the average rival distance, the average number of markets a rival serves, and the number of rival carriers.

The results for demand estimation are reported in Table H2. In line with the previous literature, we find that consumers prefer a higher share of nonstop flights and incur disutility from more miles traveled. There is strong within-nest substitution. The median own-price elasticity of -5.17 matches the literature well.

Pre-merger Supply Estimation: We consider two models of conduct for our supply specifications: (i) Bertrand pricing with constant marginal costs and (ii) a flexible supply function as described in Section 3. In this section, we focus on the flexible specification and relegate details of the Bertrand specification to Appendix H.3.

We estimate the flexible supply model with the VMM estimator described in Section 4. We include in the supply function market shares and the average distance in thousands of miles as an observable cost shifter. We also include origin-destination fixed effects. We instrument the endogenous market shares with BLP instruments formed with the following characteristics: average rival distance, average number of markets a rival serves, and number of rival carriers. Additionally, we include own-product characteristics that do not directly impact marginal costs – the share of nonstop flights and squared average distance in thousands of miles – as excluded instruments. We stratify the pre-merger data by market and split it into 80% of markets for training, leaving the remaining 20% of markets as a hold-out test sample to evaluate the fit.

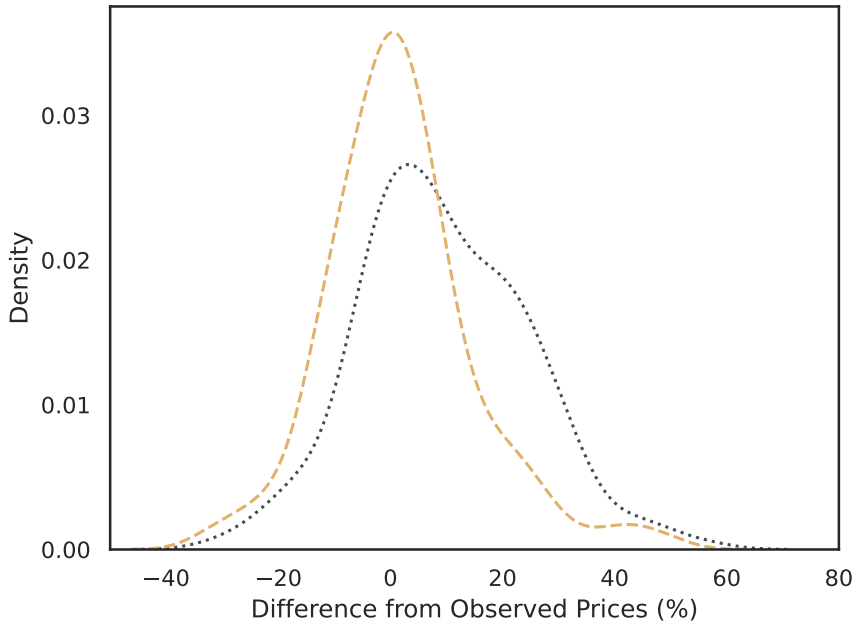
Having estimated the two models, we can evaluate their ability to explain the variation in prices in the pre-merger data. To do so, we compare the variance of the model implied estimates of the unobservable cost shock. The flexible supply function estimated with VMM significantly outperforms the parametric supply model as it implies a variance in ω that is 44% smaller than the variance implied by Bertrand conduct with constant marginal cost. This margin is similar in the training data and a test sample.

6.3 Merger Simulation Results

We examine the merger of American Airlines and US Airways in our counterfactuals. We focus on markets with three firms in the pre-merger period and two firms in the post-merger

period. We compare the predictions of the models to the true post-merger prices. The observed price differences are presented in Figure H1.

FIGURE 6: Merger Simulation Results



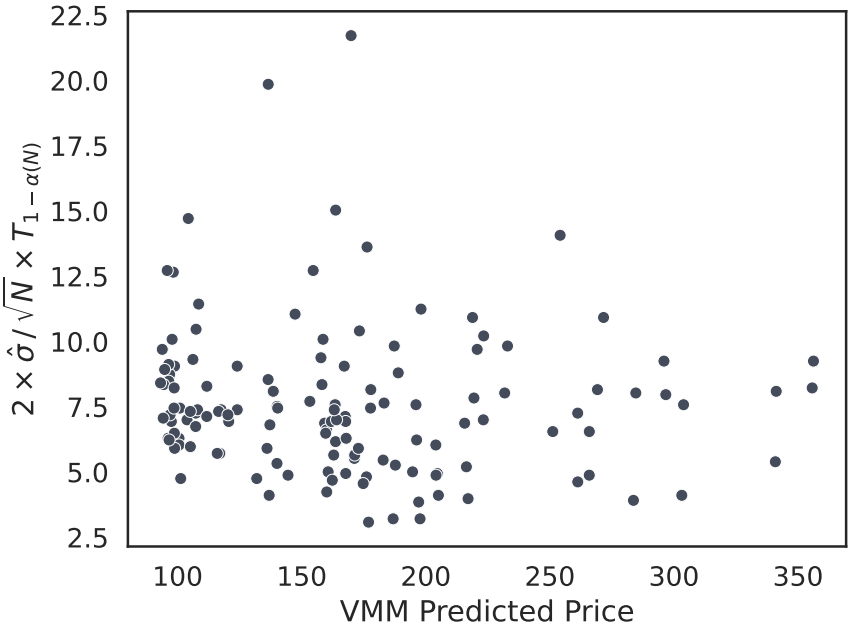
Notes: The figure reports merger simulation results for the flexible model estimated with VMM (in yellow) and the standard merger simulation model (in blue).

Figure 6 plots the distribution of percentage differences between predicted and observed prices for the two methods—the flexible supply model and the Bertrand-Nash with constant marginal cost model—in the post-merger period. The flexible model estimated with VMM is centered at zero, with a large fraction of predicted prices falling within 20% of realized prices, and a passenger-weighted MSE of 144.06. Instead, the standard merger simulation method systematically over-predicts changes and has an overall passenger-weighted MSE of 817.75. In sum, our flexible model substantially outperforms the standard simulation toolkit when predicting post-merger prices of the American-US Airways merger.

Quantifying Uncertainty: Finally, we quantify the uncertainty of our predictions in the merger simulation exercise. We follow Algorithm 1 to construct the confidence intervals. Notably, we use a more conservative Bonferroni correction for ease of exposition, allowing us to present the results with a single set of bounds for each point. We construct bounds for all points in the sample selected for merger simulation.

The results are presented in Figure 7. We present the total width of the confidence interval as a summary of uncertainty. The predictions of the inference exercise show that

FIGURE 7: Width of Confidence Intervals



Notes: The figure represents the scatter plot of VMM predicted prices (on the horizontal axis) against the width of the confidence interval for the predictions. Variance is estimated for all firms in markets affected by the US Airways-American merger.

uncertainty is roughly constant with prediction levels. Further, and more importantly, the width is at a reasonably small level in which we can make precise point estimates even with a high-dimensional markup function.

7 Conclusion

This paper demonstrates how ML/AI methods can enhance market counterfactual analysis while maintaining nonparametric economic structure. We develop a flexible supply function that nests standard oligopoly models without imposing specific assumptions about conduct or costs, enabling researchers to let the data reveal how firms actually compete. Our adaptation of neural VMM (Bennett and Kallus, 2023) provides a practical solution to the curse of dimensionality that has limited nonparametric approaches in IO, while our inference procedures deliver useful uncertainty quantification, important for policy evaluation.[23]

The Monte Carlo evidence establishes that our approach works across a few dimensions that are important for applied work: predictive accuracy even with limited sample sizes,

[23]We complete the paper with a robust framework in which researchers can easily use our method. Appendix I presents a stylized example.

scalability to realistic market sizes, successful extrapolation for policy counterfactuals, and reliable quantification of uncertainty. Our application to airline mergers demonstrates that these advantages can matter in practice: our flexible model delivers a five-fold improvement in prediction accuracy over standard assumptions. More broadly, our results suggest that misspecified conduct assumptions, which are common in applied work, may systematically bias policy recommendations.

Several limitations merit emphasis. The method requires richer instrumental variation than parametric approaches (Magnolfi and Sullivan, 2022), making it most suitable when markets exhibit substantial variation in structure and competitive conditions. While our simulations show some extrapolation ability, predictions for truly novel market configurations (such as monopolization of previously competitive markets) should be interpreted cautiously. The baseline method does not separately identify markups and costs, though we outline feasible extensions for researchers requiring further decomposition. These trade-offs make our approach as complementary to, rather than a replacement for, existing methods in the IO toolkit, including testing approaches that select among non-nested parametric models (Backus et al., 2021; Duarte et al., 2024; Dearing et al., 2024).

Our results point to new directions for integrating ML/AI with economic structure. Similar principles could extend to dynamic settings with entry and exit, to markets with more complex strategic considerations like capacity constraints or network effects, or to other economic domains where theory provides nonparametric structure and identification is based on instruments. As data availability continues to expand and computational methods advance, we anticipate growing opportunities to combine the flexibility of machine learning with the interpretability of economic models. The key insight from our analysis, that neural networks can learn economic primitives from data while respecting equilibrium constraints, suggests a path forward for empirical economics that leverages the strengths of both approaches.

References

Aryal, G., C. Murry, and J. Williams (2024): "Price Discrimination in International Airline Markets," *Review of Economic Studies*, 91, 641–689.

Azar, J., M. Schmalz, and I. Tecu (2018): "Anticompetitive Effects of Common Ownership," *Journal of Finance*, 73, 1513–1565.

Backus, M., C. Conlon, and M. Sinkinson (2021): "Common Ownership and Competition in the Ready-To-Eat Cereal Industry," NBER working paper #28350.

Bauer, B. and M. Kohler (2019): "On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression," *Annals of Statistics*, 47, 2261–2285.

Bennett, A. and N. Kallus (2023): "The Variational Method of Moments," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85, 810–841.

Berry, S. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 889–917.

Berry, S., M. Carnall, and P. Spiller (2006): "Airline Hubs: Costs, Markups and the Implications of Customer Heterogeneity," *Competition Policy and Antitrust*.

Berry, S. and P. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Berry, S. and P. Jia (2010): "Tracing the Woes: An Empirical Analysis of the Airline Industry," *American Economic Journal: Microeconomics*, 2, 1–43.

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

——— (1999): "Voluntary Export Restraints on Automobiles: Evaluating a Trade Policy," *American Economic Review*, 89, 400–431.

Betancourt, J., A. Hortaçsu, A. Oery, and K. Williams (2022): "Dynamic Price Competition: Theory and Evidence from Airline Markets," NBER Working Paper #30347.

Björnerstedt, J. and F. Verboven (2016): "Does Merger Simulation Work? Evidence from the Swedish Analgesics Market," *American Economic Journal: Applied Economics*, 8, 125–164.

Bontemps, C., C. Gualdani, and K. Remmy (2023): "Price Competition and Endogenous Product Choice in Networks: Evidence from the US Airline Industry," Working Paper.

Brand, J. and A. Smith (2025): "A Quasi-Bayes Approach to Nonparametric Demand Estimation with Economic Constraints," *Available at SSRN 5100826*.

Cardell, N. (1997): "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 13, 185–213.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization," *Handbook of econometrics*, 6, 5633–5751.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Amsterdam: Elsevier, vol. 6, chap. 76, 5549–5632.

Ciliberto, F., C. Murry, and E. Tamer (2021): "Market Structure and Competition in Airline Markets," *Journal of Political Economy*, 129, 2995–3038.

Ciliberto, F. and J. Williams (2014): "Does Multimarket Contact Facilitate Tacit Collusion? Inference on Conduct Parameters in the Airline Industry," *RAND Journal of Economics*, 45, 764–791.

Compiani, G. (2022): "Market Counterfactuals and the Specification of Multiproduct Demand: A Nonparametric Approach," *Quantitative Economics*, 13, 545–591.

Dearing, A., L. Magnolfi, D. Quint, C. Sullivan, and S. Waldfogel (2024): "Learning Firm Conduct: Pass-through as a Foundation for Instrument Relevance," Working paper.

Dikkala, N., G. Lewis, L. Mackey, and V. Syrgkanis (2020): "Minimax Estimation of Conditional Moment Models," *Advances in Neural Information Processing Systems*, 33, 12248–12262.

Duarte, M., L. Magnolfi, M. Sølvsten, and C. Sullivan (2024): "Testing Firm Conduct," *Quantitative Economics*, 15, 571–606.

Dubois, P., R. Griffith, and M. O'Connell (2018): "The Effects of Banning Advertising in Junk Food Markets," *Review of Economic Studies*, 85, 396–436.

Farrell, M., T. Liang, and S. Misra (2020): "Deep Learning for Individual Heterogeneity: An Automatic Inference Framework," *arXiv preprint arXiv:2010.14694*.

Gandhi, A. and J.-F. Houde (2020a): "Measuring Firm Conduct in Differentiated Products Industries," Working Paper.

——— (2020b): "Measuring Substitution Patterns in Differentiated Products Industries," Working Paper.

Goldberg, P. (1998): "The Effects of the Corporate Average Fuel Efficiency Standards in the US," *Journal of Industrial Economics*, 46, 1–33.

Gowrisankaran, G., A. Nevo, and R. Town (2015): "Mergers when Prices are Negotiated: Evidence from the Hospital Industry," *American Economic Review*, 105, 172–203.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017): "Deep IV: A Flexible Approach for Counterfactual Prediction," in *International Conference on Machine Learning*, PMLR, 1414–1423.

Holm, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Hong, H., A. Mahajan, and D. Nekipelov (2015): "Extremum Estimation and Numerical Derivatives," *Journal of Econometrics*, 188, 250–263.

Hortaçsu, A., O. Natan, H. Parsley, T. Schwieg, and K. Williams (2024): "Organizational Structure and Pricing: Evidence from a Large US Airline," *Quarterly Journal of Economics*, 139, 1149–1199.

Ito, K. and J. Sallee (2018): "The Economics of Attribute-based Regulation: Theory and Evidence from Fuel Economy Standards," *Review of Economics and Statistics*, 100, 319–336.

Jacobsen, M., J. Sallee, J. Shapiro, and A. Van Benthem (2023): "Regulating Untaxable Externalities: Are Vehicle Air Pollution Standards Effective and Efficient?" *Quarterly Journal of Economics*, 138, 1907–1976.

Kaji, T., E. Manresa, and G. Pouliot (2023): "An Adversarial Approach to Structural Estimation," *Econometrica*, 91, 2041–2063.

Kennedy, P., D. O'Brien, M. Song, and K. Waehrer (2017): "The Competitive Effects of Common Ownership: Economic Foundations and Empirical Evidence," *Available at SSRN 3008331*.

Lewis, G. and V. Syrgkanis (2018): "Adversarial Generalized Method of Moments," *arXiv preprint arXiv:1803.07164*.

Li, S., J. Mazur, Y. Park, J. Roberts, A. Sweeting, and J. Zhang (2022): "Repositioning and Market Power after Airline Mergers," *RAND Journal of Economics*, 53, 166–199.

Liao, L., Y.-L. Chen, Z. Yang, B. Dai, M. Kolar, and Z. Wang (2020): "Provably Efficient Neural Estimation of Structural Equation Models: An Adversarial Approach," *Advances in Neural Information Processing Systems*, 33, 8947–8958.

Magnolfi, L., D. Quint, C. Sullivan, and S. Waldfogel (2022): "Differentiated-Products Cournot Attributes Higher Markups than Bertrand–Nash," *Economics Letters*, 219, 110804.

Magnolfi, L. and C. Sullivan (2022): "A Comparison of Testing and Estimation of Firm Conduct," *Economics Letters*, 212, 110316.

Miller, N. and M. Weinberg (2017): "Understanding the Price Effects of the MillerCoors Joint Venture," *Econometrica*, 85, 1763–1791.

Miravete, E., K. Seim, and J. Thurk (2018): "Market Power and the Laffer Curve," *Econometrica*, 86, 1651–1687.

Morrow, W. and S. Skerlos (2011): "Fixed-Point Approaches to Computing Bertrand-Nash Equilibrium Prices Under Mixed-Logit Demand," *Operations Research*, 59, 328–345.

Nevo, A. (2000): "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *RAND Journal of Economics*, 395–421.

Newey, W. and J. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Otsu, T. and M. Pesendorfer (2024): "Conduct in the Soft Drink Market: A Mechanism Design Approach," Working Paper.

Peters, C. (2006): "Evaluating the Performance of Merger Simulation: Evidence from the US Airline Industry," *Journal of Law and Economics*, 49, 627–649.

Petrin, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110, 705–729.

Schmidt-Hieber, J. (2020): "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function," *Annals of Statistics*, 48, 1875–1897.

Sill, J. (1997): "Monotonic networks," *Advances in Neural Information Processing Systems*, 10, 661–667.

Small, K. and H. Rosen (1981): "Applied Welfare Economics with Discrete Choice Models," *Econometrica*, 105–130.

Tebaldi, P., A. Torgovitsky, and H. Yang (2023): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," *Econometrica*, 91, 107–146.

Wehenkel, A. and G. Louppe (2019): "Unconstrained monotonic neural networks," *Advances in neural information processing systems*, 32.

Wei, Y. and Z. Jiang (2025): "Estimating Parameters of Structural Models Using Neural Networks," *Marketing Science*, 44, 1–246.

Werden, G. and L. Froeb (1994): "The Effects of Mergers in Differentiated Products Industries: Logit Demand and Merger Policy," *Journal of Law, Economics, & Organization*, 10, 407–426.

Williams, K. (2022): "The Welfare Effects of Dynamic Pricing: Evidence from Airline Markets," *Econometrica*, 90, 831–858.

You, S., D. Ding, K. Canini, J. Pfeifer, and M. Gupta (2017): "Deep Lattice Networks and Partial Monotonic Functions," *Advances in Neural Information Processing Systems*, 30.

Yuan, Z. and P. J. Barwick (2024): "Network Competition in the Airline Industry: An Empirical Framework," Tech. rep., National Bureau of Economic Research.

Zhang, R., M. Imaizumi, B. Schölkopf, and K. Muandet (2023): "Instrumental Variable Regression via Kernel Maximum Moment Loss," *Journal of Causal Inference*, 11, 20220073.

A Extensions: Decomposing Costs and Markups, and Other Economic Restrictions

While our main identification result in Section 3.2 establishes nonparametric identification of the supply function $h_j(s_t, D_t, \mathrm{w}_{jt})$, certain counterfactual exercises require separate identification of the cost and markup components. For instance, evaluating technological improvements requires knowledge of the cost function, while assessing changes in competitive conduct requires identification of markups. We outline two approaches to achieve this decomposition, though we do not pursue these extensions because they lie outside our primary scope.

A.1 Separating Markup and Cost Through Market Size Variation

The first approach leverages exogenous variation in market size M_t to separately identify cost and markup functions without imposing parametric restrictions on either component.

Recall that quantities are $q_{jt} = M_t s_{jt}$. When market size varies, holding market shares fixed, costs change through the quantity channel while markups remain constant because they depend only on shares and demand derivatives. Under the decomposition:

$$h_j(s_t, D_t, \mathrm{w}_{jt}) = c_j(M_t s_{jt}, \mathrm{w}_{jt}) + \Delta_j(s_t, D_t),$$

the key insight is that for fixed (s, D, w), variation in M_t affects only the cost component. Taking the derivative with respect to M:

$$\frac{\partial}{\partial M}\mathbb{E}\left[p_{jt} \mid s_t = s, D_t = D, \mathrm{w}_{jt} = \mathrm{w}, M_t = M\right] = \frac{\partial c_j}{\partial q_{jt}} \cdot s.$$

This identifies the marginal cost at each quantity level. Integrating from a baseline quantity q_0:

$$c_j(q, \mathrm{w}) - c_j(q_0, \mathrm{w}) = \int_{q_0}^{q} \frac{1}{s}\frac{\partial}{\partial M}\mathbb{E}\left[p_{jt} \mid s_t = s, D_t = D, \mathrm{w}_{jt} = \mathrm{w}, M_t = \tilde{q}/s\right] d\tilde{q}.$$

With a normalization for $c_j(q_0, \mathrm{w})$, the cost function is identified. The markup function follows as:

$$\Delta_j(s, D) = \mathbb{E}\left[p_{jt} \mid s_t = s, D_t = D, \mathrm{w}_{jt} = \mathrm{w}, M_t = M\right] - c_j(Ms, \mathrm{w}).$$

This approach requires substantially stronger data requirements than our baseline identification. Most importantly, it requires observing the same market configuration (s_t, D_t) across different market sizes—essentially requiring that M_t varies independently of equilibrium shares and demand derivatives. This may be violated if market size itself affects equilibrium outcomes through entry, product positioning, or competitive intensity. Moreover, the completeness condition must be augmented to ensure that variation in M_t provides sufficient information to separately identify both

cost and markup functions. In many empirical settings, such rich variation in market size may be unavailable or confounded with other market characteristics.

A.2 Separating Markup and Cost Through Economic Restrictions

The second approach imposes structure on either conduct or costs to achieve separate identification. We consider each possibility in turn.

A.2.1 Known Conduct

Following Berry and Haile (2014) Section 4.3, suppose the form of oligopoly competition is known:

Assumption 9 (Known Oligopoly Model)**.** Markups take the form $\Delta_j(s_t, D_t) = \psi_j(s_t, D_t)$ where ψ_j are known functions determined by the oligopoly model.

For example, under Bertrand-Nash competition with multi-product firms, ψ_j is the (j,j)-th element of the matrix Γ_t^{-1} where the (j,k)-th element of Γ_t is equal to $\partial s_k/\partial p_j$ when products j and k are produced by the same firm, and zero otherwise.

Under Assumption 9, marginal costs are directly identified:

$$c_j(q_{jt}, \mathrm{w}_{jt}) + \omega_{jt} = p_{jt} - \psi_j(s_t, D_t).$$

The cost function $c_j(\cdot,\cdot)$ can then be identified nonparametrically using instruments for quantity, following the arguments in Berry and Haile (2014) Theorem 6. This requires that demand shifters x_{jt} are excluded from marginal costs, allowing them to serve as instruments for the endogenous quantity in the cost function regression.

The key advantage of this approach is that it requires only standard instrumental variables variation rather than the stringent market size conditions of Section A.1. The disadvantage is the strong assumption of known conduct, which rules out testing between alternative models of competition.

A.2.2 Parametric Costs

Alternatively, we can impose a functional form for costs while maintaining flexibility in conduct:

Assumption 10 (Parametric Cost Structure)**.** Marginal costs take the parametric form $c_j(q_{jt}, \mathrm{w}_{jt}; \theta_j) = c(\mathrm{w}_{jt}, q_{jt}; \theta_j)$ for a known function $c(\cdot,\cdot;\theta_j)$ with finite-dimensional parameters θ_j.

A common specification is log-linear costs: $c_j(q_{jt}, \mathrm{w}_{jt}; \theta_j) = \mathrm{w}_{jt}'\gamma_j + \alpha_j \log(q_{jt})$ where $\theta_j = (\gamma_j, \alpha_j)$. This transforms our nonparametric identification problem into a semi-parametric one. The supply equation becomes:

$$p_{jt} = c(\mathrm{w}_{jt}, M_t s_{jt}; \theta_j) + \Delta_j(s_t, D_t) + \omega_{jt}.$$

The parameters θ_j can be estimated in a first stage using the orthogonality condition $\mathbb{E}[\omega_{jt}|z_t, \mathrm{w}_t] = 0$ and excluded instruments. Given consistent estimates $\hat{\theta}_j$, the markup function is identified non-parametrically as:

$$\Delta_j(s_t, D_t) = h_j(s_t, D_t, \mathrm{w}_{jt}) - c(\mathrm{w}_{jt}, M_t s_{jt}; \hat{\theta}_j)$$

This semi-parametric approach represents a middle ground: it imposes less structure than assuming known conduct while requiring weaker data requirements than the fully nonparametric approach of Section A.1. The parametric restrictions on costs are often more palatable than conduct assumptions, as they can be motivated by production theory or tested against more flexible specifications.

A.3 Implementing Economic and Statistical Restrictions

As discussed above, one may want to impose additional economic or statistical restrictions on the h function. In addition to the separability between cost and markups, one may want, e.g., h_j to be decreasing in that product's own demand elasticity, as is the case in many standard models. This is in the spirit of the micro-founded economic restrictions that are imposed on nonparametric demand systems in Compiani (2022) and Brand and Smith (2025). It is possible to add these restrictions to our model through regularization, restrictions on weights and activation functions, neural network architecture, or some combination of these. We discuss each of these in turn within the example of monotonicity in own demand elasticities.

The first approach is to incorporate additional components in the regularization term R_N in Equation (4). Define a set of n own-demand elasticities in increasing order as $D = (D_{(1)}, ..., D_{(n)})$. An example of a simple regularization term R^M for monotonicity is the following:

$$R^M(h) = \sum_{i=2}^{n} (\max\{h(D_{(i)}) - h(D_{(i-1)}), 0\})^2$$

We condition on s_t, w_t, θ, and $\mathcal{H}_t$ in the markup function h, suppressing them for notational simplicity. If $h(D_{(i)}) \leq h(D_{(i-1)})$, there is no additional penalty on the markup function, but otherwise, we penalize the squared first difference of own-demand elasticities in the spirit of a ridge regression. We note that the choice of regularization is a degree of freedom for the researcher.

The last two approaches are related to the rich computer science literature on monotonic neural networks, starting with Sill (1997). The first approach enforces constraints on weights and activation functions in particular neural network layers, e.g., You, Ding, Canini, Pfeifer, and Gupta (2017). The user can specify inputs, such as own-demand elasticities, in which the output is monotonic. A second approach, detailed in Wehenkel and Louppe (2019), uses the architecture of the neural network to enforce a constant sign of the derivative of the approximated function without imposing additional constraints. In our specific example, we can restrict the derivative of h with respect to own-demand elasticity to be negative.

The discussion above focuses on a single example of an economic restriction. These approaches can be adapted and combined to impose additional economic or statistical restrictions on the markup function.

B Identification of the Supply Function: Proofs and Additional Discussion

B.1 Proof of Theorem 2

The proof adapts the standard NPIV logic from Newey and Powell (2003) to our manifold setting. Under Assumptions 1-7, suppose two functions h_j^1 and h_j^2 both satisfy the moment condition in Equation (3). Then:

$$\mathbb{E}[(h_j^1 - h_j^2)(s_t, D_t, \mathrm{w}_{jt}) \mid z_{jt}, \mathrm{w}_{jt}] = 0.$$

By the manifold completeness condition (Assumption 7), this implies $(h_j^1 - h_j^2) = 0$ almost surely on $\mathcal{M}$. Therefore, h_j is uniquely identified on the manifold. This follows directly from Proposition 2.1 in Newey and Powell (2003), applied to the manifold $\mathcal{M}$ rather than the full space.

B.2 Understanding Manifold Completeness

The manifold completeness condition requires that instrumental variation be rich enough to identify the supply function on all feasible combinations of (s_t, D_t). To understand why this condition is reasonable, consider the necessary rank condition that makes completeness meaningful.

Given the known demand system and index structure, the manifold $\mathcal{M}$ is implicitly defined by the constraint $D_t = D(s_t, \delta_t, x_t^{(2)})$ where $\delta_t = x_t^{(1)} + \xi_t$. For identification on this manifold, we need the rank condition:

$$\operatorname{rank}\left[\frac{\partial \operatorname{vec}(D(s_t, \delta_t, x_t^{(2)}))}{\partial(\delta_t, \operatorname{vec}(x_t^{(2)}))}\right] = JK.$$

This rank condition ensures that variations in the demand index δ_t and characteristics $x_t^{(2)}$ generate sufficient independent directions of movement in the demand derivatives. Without this property, different supply functions could be observationally equivalent, making completeness vacuous. The rank condition is not a separate assumption but rather clarifies what we mean by completeness in this setting. It guarantees that variation in $x_t^{(1)}$ shifts the demand index $\delta_t = x_t^{(1)} + \xi_t$, variation in $x_t^{(2)}$ directly affects demand, and together with rival cost shifters $\mathrm{w}_{-j,t}$ affecting equilibrium shares, these instruments span the manifold.

The key insight is dimensional reduction through the manifold structure. While (s_t, D_t) live in $\mathbb{R}^{J+J^2}$, the true endogenous variation comes only from (s_t, δ_t), which has dimension $2J$. The exogenous characteristics $x_t^{(2)}$ can be conditioned upon, leaving only $2J$ dimensions of endogenous

variation that need to be instrumented. This represents a dramatic reduction from the $J + J^2$ dimensions of the full space. For example, with $J = 30$ products, standard completeness would require instruments for variation in $30 + 900 = 930$ dimensions, while our approach only needs to handle 60 dimensions of endogenous variation coming from (s_t, δ_t). This dimensional reduction, from $J+J^2$ to $2J$, makes nonparametric identification feasible even in markets with many products.

B.3 Comparison with Berry and Haile (2014)

Our identification strategy is related to results in Section 4.4 in Berry and Haile (2014), who identify the inverse supply function $\pi^{-1}(s_t, p_t)$ mapping market shares and prices to values of a cost index $\kappa_{jt} = \mathrm{w}_{jt}^{(1)} + \omega_{jt}$.

The approaches share several key features. Both use exclusion restrictions, where some product characteristics are excluded from costs, and both require completeness conditions, although on different spaces. However, we identify $h_j(s_t, D_t, \mathrm{w}_{jt}) = \Delta_j + c_j$ directly, while Berry and Haile (2014) identify the inverse π^{-1}. While both formulations of supply may be used for some counterfactuals, our direct estimation of the supply function lends itself to further decomposition into markups and cost, an approach that we discuss in Appendix A.1.

C Data Construction for Ownership-Based Ordering

This appendix details the implementation of the ownership-based ordering described in Assumption 8. We show how market shares, demand derivatives, and instruments are systematically reordered to embed ownership structure into the supply function estimation.

C.1 General Construction Procedure

Consider a market t with products indexed by j and firms indexed by f. For each product j owned by firm f, we construct the following ordered vectors:

Outcomes: The dependent variable is simply the price p_{jt}.

Exogenous Variables: Own cost shifters w_{jt} enter directly without reordering.

Market Shares: We reorder market shares using the ownership matrix. First, we place the own market share s_{jt}. Second, we include other products owned by the same firm, denoted $s_{-j,f,t}$. If firm f owns fewer than the maximum number of products $\bar{J}_f$ observed across all firms in the sample, we pad with zeros to maintain consistent dimensions. Third, we append rival firms' market shares $s_{-f,t}$.

Demand Derivatives: The demand derivative matrix is partitioned into blocks following the ownership structure. We order own-price elasticities as $(D_{jj,t}, D_{(-j,f),(-j,f),t}, D_{-f,-f,t})$, padding with zeros as needed. Cross-price elasticities are ordered as own-firm cross-elasticities $D_{j,(-j,f),t}$ followed by cross-firm elasticities $D_{j,-f,t}$.

Instruments: Product characteristics and rival cost shifters are ordered following the same pattern as market shares: $(x_{jt}, x_{-j,f,t}, x_{-f,t}, \mathrm{w}_{-f,t})$.

C.2 Example: Duopoly with Multi-Product Firm

Consider a market with a single-product firm (Firm 0) and a two-product firm (Firm 1). The raw data contains:

Table C1: Raw Product Data

Firm	Product	x_1	w_1	s	p
0	0	0.4	0.1	0.4	2
1	1	0.2	0.2	0.2	3
1	2	0.3	0.3	0.3	4

Under logit demand with price coefficient $\alpha = -1$, the demand derivative matrix is:

$$D_t = \begin{bmatrix} -0.24 & 0.08 & 0.12 \\ 0.08 & -0.16 & 0.06 \\ 0.12 & 0.06 & -0.21 \end{bmatrix}$$

After applying the ownership-based ordering, the constructed data for estimation becomes:

Table C2: Reordered Market Shares and Cost Shifters

			Own Firm		Rival Firm		
Firm	Product	w_1	s_j	$s_{-j,f}$	$s_{-f,0}$	$s_{-f,1}$	p
0	0	0.1	0.4	0	0.2	0.3	2
1	1	0.2	0.2	0.3	0.4	0	3
1	2	0.3	0.3	0.2	0.4	0	4

Note that Product 0, being the only product of Firm 0, has $s_{-j,f} = 0$ (padded). Products 1 and 2, both owned by Firm 1, include each other's shares in the $s_{-j,f}$ column and pad the second rival column with zeros since only one rival firm exists.

The demand derivatives are similarly reordered:

TABLE C3: Reordered Own-Price Demand Derivatives

		Own Elasticities		Rival Own-Elasticities	
Firm	Prod.	D_{jj}	$D_{(-j,-j),f}$	$D_{-f0,-f0}$	$D_{-f1,-f1}$
0	0	-0.24	0	-0.16	-0.21
1	1	-0.16	-0.21	-0.24	0
1	2	-0.21	-0.16	-0.24	0

TABLE C4: Reordered Cross-Price Demand Derivatives

		j Own Cross	j Rival Cross		$-j$ Own Cross	$-j$ Rival Cross
Firm	Prod.	$D_{j,f}$	$D_{j,-f0}$	$D_{j,-f1}$	$D_{-j,f}$	$D_{(-j,-f),-f}$
0	0	0	0.08	0.12	0	0.06
1	1	0.06	0.08	0	0.12	0
1	2	0.06	0.12	0	0.08	0

This systematic reordering allows the neural network to learn patterns that depend on ownership structure without explicitly including the ownership matrix as a separate input, effectively embedding the economic structure of multi-product firm decision-making into the functional form of the estimator.

D On Deep Neural Networks Architectures

In a nonparametric regression framework, Schmidt-Hieber (2020) demonstrates that estimators based on sparsely connected deep neural networks (DNNs) with ReLU activation functions attain the minimax rate of convergence. In Section 5, the author further shows that nonparametric regression using wavelet bases achieves only suboptimal rates. Drawing on these insights from Schmidt-Hieber (2020), we provide an explanation for why our DNN-based VMM estimator outperforms its series-based NPIV counterpart.

Let $\sigma(x) = \max\{0, x\}$. For $\mathbf{v} = (v_1, ..., v_r) \in \mathbb{R}^r$, $r \in \mathbb{N}$, define:

$$\sigma_{\mathbf{v}}(y_1, ..., y_r) = (\sigma(y_1 - v_1), ..., \sigma(y_r - v_r)).$$

A neural network with network architecture $(L, \mathbf{p})$, $L > 0$ the number of hidden layer, $\mathbf{p} =$

$(p_1, ..., p_{L+1}) \in \mathbb{N}^{L+2}$ a width vector, is a function of the form:

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad [0,1]^d \ni \mathbf{x} \mapsto f(\mathbf{x}) = W_L \sigma_{\mathbf{v}_L} W_{L-1} \sigma_{\mathbf{v}_{L-1}} ... W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x}, \tag{D1}$$

where W_i is a $p_{i+1} \times p_i$ weight matrix and $\mathbf{v}_i \in \mathbb{R}^{p_i}$. Define the set of s-sparse networks as:

$$\mathcal{F}(L, \mathbf{p}, s, F) = \left\{ f \text{ of the form (D1)} : \max_{j=0,...,L} \|W_j\|_\infty \vee |\mathbf{v}_j|_\infty \leq 1, \sum_{j=0}^{L} \|W_j\|_0 + |\mathbf{v}_j|_0 \leq s, \| |f|_\infty \|_\infty \leq F \right\}.$$

Also define the set of functions that can be expressed as compositions of some Hölder functions by:

$$\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \beta, K) = \Big\{ f = g_q \circ \cdots \circ g_0 : g_\ell = (g_{\ell m})_m : [a_i, b_i]^{d_\ell} \to [a_{\ell+1}, b_{\ell+1}]^{d_{\ell+1}},$$
$$g_{\ell m} \in C^{\beta_\ell}_{t_\ell}([a_\ell, b_\ell]^{t_\ell}, K) \text{ for some } |a_\ell|, |b_\ell| \leq K \Big\},$$

where $C^\beta_r(D, K)$ denotes the set of real-valued functions defined on a domain $D \subset \mathbb{R}^r$ that belong to the Hölder space of smoothness β and have Hölder norm bounded by K. In the Proof of Theorem 1 in Schmidt-Hieber (2020), it was shown in Equation (26) and the following paragraph that:

$$\inf_{f^* \in \mathcal{F}(L, \mathbf{p}, s, F)} \|f^* - f_0\|^2_\infty \leq C \max_{\ell=0,...,q} c^{-\frac{2\beta^*_\ell}{t_\ell}} n^{-\frac{2\beta^*_\ell}{2\beta^*_\ell + t_\ell}} \tag{D2}$$

for the function $f_0 \in \mathcal{G}(q, \mathbf{d}, \mathbf{t}, \beta, K)$ and a constant $C > 0$, where t_ℓ is the maximal number of variables on which each of the $g_{\ell m}$ depends on. It is important to note that this is a purely approximation-theoretic result; no assumptions are made regarding the structure of the error term in the regression model. Moreover, the ambient dimension d of the function f_0 does not appear directly in the convergence bound. Instead, the rate depends on the intrinsic dimensions t_ℓ of the latent functions within the compositional structure of f_0.

We observe that each component of the markup function $f_0(\cdot)$ arising from Bertrand-Nash competition can be well approximated by a deep neural network (DNN), as it belongs to the function class $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$. To see this, note that following Section 2.2 of Magnolfi, Quint, Sullivan, and Waldfogel (2022), the j-th component of the (recentered) Bertrand–Nash markup function can be written as:

$$f_{0j}(\mathbf{p}, \mathbf{c}, \Omega, S_p) = p_j - c_j - \left[(\Omega \odot S'_p)^{-1} s \right]_j,$$

where $\mathbf{p}$ and $\mathbf{c}$ are vectors of prices and marginal costs, respectively, Ω is the ownership matrix, and S'_p denotes the matrix of price derivatives of market shares. Typically, Ω exhibits a sparse block structure, which implies that $(\Omega \odot S'_p)^{-1}$ also inherits sparsity. This structural sparsity

corresponds to a form of sparse tensor decomposition, as considered in Equation (14) of Schmidt-Hieber (2020), and hence places f_0 within the compositional Hölder class $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$. The implication is that DNNs can approximate such functions at the faster rate given in Equation (D2), which depends on the latent dimensions t_ℓ, rather than the ambient dimensionality d typically governing nonparametric approximation rates.

Although series-based estimators also rely on the idea that an unknown function can be approximated by a linear combination of a finite number of basis functions, they require an explicit choice of basis. This choice can significantly influence the quality of approximation and, consequently, the convergence rate—unlike in the case of DNN, which are data-adaptive and less reliant on prior structural assumptions.

Wavelet series, for instance, are well studied and known to possess attractive theoretical properties. However, (Schmidt-Hieber, 2020, Section 5) shows that even for relatively simple target functions—such as additive models of the form:

$$f_0(\mathbf{x}) = h(x_1 + \cdots + x_d), \quad \text{with } h \in C_1^\alpha([0, d], K),$$

series approximations using wavelet bases may still attain suboptimal rates. Specifically, in the proof of Theorem 4, it is shown that for any $0 < \alpha \leq 1$ and Hölder radius $K > 0$, the following lower bound on the approximation error holds:

$$\sup_{h \in C_1^\alpha([0,d],K)} \|f^* - f_0\|_\infty^2 \geq C n^{-2\alpha/(2\alpha+d)},$$

for some constant $C > 0$, f^* is any function constructed from compactly supported wavelet bases. We anticipate that a similar result can be established for the more nonlinear structures present in the current formulation of the markup function.

Recent advances in approximation theory highlight the advantages of DNNs over traditional series-based estimators in nonparametric settings. Unlike series-based methods, which rely on a pre-specified basis and may suffer from suboptimal convergence, DNNs adapt to the structure of the target function and can achieve faster approximation rates. This advantage is especially relevant for structural models such as those arising from supply side competition, where the markup function exhibits compositional and sparse structures well suited to DNN approximation. These insights suggest that DNN-based VMM-estimator extend their benefits to more complex, nonlinear formulations of the markup function.

E Omitted Details for Quantification of Uncertainty

The following theorem provides the main theoretical foundation for our inference procedure. Following Bennett and Kallus (2023), we assume throughout this section that the true parameter value

θ_0 lies in Θ, a compact subset of a finite-dimensional Euclidean space.[24]

Theorem 7. *Let $\tilde{\theta}_N \xrightarrow{p} \theta_0$. Suppose that $\Theta \ni h \mapsto \mathbb{R}^d$ is differentiable at $\theta_0 \in \Theta$ for each given $\mathcal{H}$. Under regularity conditions imposed by Theorems 2-3 in Bennett and Kallus (2023), we have:*

$$\left\{\nabla_{\theta'} h(\theta_0, \mathcal{H}) \Omega_0^{-1} \nabla_{\theta'} h(\theta_0, \mathcal{H})'/N\right\}^{-1/2} (h(\hat{\theta}_N, \mathcal{H}) - h(\theta_0, \mathcal{H})) \xrightarrow{d} N(0, I). \tag{E1}$$

Proof. Suppose $\tilde{\theta}_N \xrightarrow{p} \theta_0$ and the regularity conditions from Theorems 2-3 in Bennett and Kallus (2023) hold. This implies:

$$\{\Omega_0/N\}^{-1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, I),$$

where $\Omega_0 = \mathbb{E}\left[\mathbb{E}[\nabla_\theta \omega(\theta_0) \mid z, w]' \mathbb{E}[\omega(\theta)\omega(\theta)' \mid z, w]^{-1} \mathbb{E}[\nabla_\theta \omega(\theta_0) \mid z, w]\right]$. Now apply the function h to the set of d observations. The result follows directly from the delta method.

□

Next, we extend the asymptotic result to cover smooth functions of prices.

Assumption 11. *(Smooth Counterfactuals)* The counterfactual map F is continuous and once differentiable in prices such that $\nabla_p F$ (or equivalently $\nabla_h F(h(\theta))$) exists.

Assumption 11 is satisfied by a large range of counterfactuals, e.g., logit shares, consumer surplus, and government revenue. We can construct confidence intervals by applying the delta method to the specific composite function $F(h(\theta))$.

Theorem 8. *Under Assumption 11 and the regularity conditions in Theorems 2-3 in Bennett and Kallus (2023), we have:*

$$\left\{\nabla_h F(h(\theta_0)) \nabla_\theta h(\theta_0) \Omega_0^{-1} \nabla_\theta h(\theta_0)' \nabla_h F(h(\theta_0))'/N\right\}^{-1/2} (F(h(\hat{\theta}_N)) - F(h(\theta_0))) \xrightarrow{d} N(0, I). \tag{E2}$$

Proof. The numerical delta method in Hong et al. (2015) implies Equation (E1). Applying the delta method under Assumption 11 to the composite function $F(h(\theta))$ to Equation (E1), we find Equation (E2).

□

Taking $\beta = \nabla_h F(h_x(\theta_0)) \nabla_\theta h_x(\theta_0)$ for some selected set of products x, Lemma 9 in Bennett and Kallus (2023) implies that the asymptotic variance for $F(h_x(\theta))$ in Equation (E2) can be estimated using the same method as described in Equation (5). We can construct simultaneous confidence intervals for many predictions, e.g., market shares, using Algorithm 1 or for singletons, e.g., total consumer surplus, directly with Equation (5). Below, we present a series of corollaries for specific counterfactuals that correspond to Examples 4-6 and more.

[24] Compactness is required only for establishing asymptotic normality, not for consistency.

Corollary 1. *Suppose we would like to conduct inference on counterfactual market share for product j with the map $F(h(\theta)) = s_j(h(\theta))$. We have:*

$$\left\{ D_j(h(\theta_0))\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)' D_j(h(\theta_0))'/N \right\}^{-1/2} (F(h(\hat{\theta}_N)) - F(h(\theta_0))) \xrightarrow{d} N(0, I),$$

where D_j is the row corresponding to product j in the demand derivative matrix evaluated at counterfactual prices $h(\theta)$.

Proof. For a generic demand system, we find that $\nabla_h F(h(\theta)) = D_j(h(\theta))$. For example, take logit demand. Using the functional form of logit, we find $[\nabla_h F(h(\theta))]_k$ at index k as:

$$[\nabla_h F(h(\theta))]_k = [D_j(h(\theta))]_k = \begin{cases} -\alpha s_j(h(\theta))(1 - s_j(h(\theta)) & j = k \\ \alpha s_j(h(\theta)) s_k(h(\theta)) & j \neq k \end{cases}$$

Stacking the elements into a $1 \times J$ matrix, this is equivalent to row j of the derivative matrix of demand evaluated at the counterfactual $h(\theta)$, defined above as $D_j(\cdot)$. Thus, more generically, $\nabla_h F(h(\theta)) = D_j(h(\theta))$. The result follows immediately from Theorem 8.

□

Corollary 2. *Denote $\mathcal{J} = \mathcal{J}_1 \cup \cdots \cup \mathcal{J}_M$ as the union of product sets across markets $m = 1, ..., M$. Suppose we would like to conduct inference on counterfactual the total quantity across markets with the map $F(h(\theta)) = \sum_{j \in \mathcal{J}} s_j(h(\theta))$. We have:*

$$\left\{ 1'_{|\mathcal{J}|} D(h(\theta_0))\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)' D(h(\theta_0))' 1_{|\mathcal{J}|}/N \right\}^{-1/2} (F(h(\hat{\theta}_N)) - F(h(\theta_0))) \xrightarrow{d} N(0, I),$$

where $\nabla_h F(h(\theta)) = 1'_{|\mathcal{J}|} D(h(\theta))$. $D(\cdot)$ is stacked block-diagonal across markets and evaluated at counterfactual prices $h(\theta)$.

Proof. Taking the gradient and borrowing notation from the previous proof, we find:

$$[\nabla_h F(h(\theta))]_j = \sum_{k \in \mathcal{J}} [D(h(\theta))]_{kj}.$$

Notice that, by market independence, $[D(h(\theta))]_{kj} = 0$ whenever j and k are in different markets. Stacking the gradients $D(\cdot)$ block-diagonal across markets, we have a $1 \times |\mathcal{J}|$ matrix $\nabla_h F(h(\theta)) = 1'_{|\mathcal{J}|} D(h(\theta))$. The result follows immediately from Theorem 8.

□

Corollary 3. *Suppose we would like to conduct inference on total consumer surplus across markets $m = 1, ..., M$ from logit demand with the map*

$$F(h(\theta)) = -\frac{1}{\alpha} \sum_m \log \left(1 + \sum_{j \in \mathcal{J}_m} \exp(-\alpha h_j(\theta) + x_j\beta + \xi_j) \right).$$

. *We have:*

$$\left\{s(h(\theta_0))'\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)'s(h(\theta_0))/N\right\}^{-1/2}(F(h(\hat{\theta}_N))-F(h(\theta_0)))\xrightarrow{d}N(0,I),$$

where $s(h(\theta))$ are market shares evaluated at counterfactual prices $h(\theta)$.

Proof. Taking the gradient of the map F with respect to h, we find the j-th element as:

$$[\nabla_h F(h(\theta))]_j=\frac{\exp(-\alpha h_j(\theta)+x_j\beta+\xi_j)}{1+\sum_{k\in\mathcal{J}_{m(j)}}\exp(-\alpha h_k(\theta)+x_k\beta+\xi_k)}=s_j(h(\theta)).$$

Notice that, by market independence, gradients with respect to prices in other markets are zero. Stacking the gradients across markets, we have a $1\times|\mathcal{J}|$ matrix of transposed market shares $s(\cdot)'$ evaluated at the counterfactual $h(\theta)$. The result follows immediately from Theorem 8.

□

Corollary 4. *Denote $\mathcal{J}=\mathcal{J}_1\cup\cdots\cup\mathcal{J}_M$ as the union of product sets across markets $m=1,...,M$. Suppose we would like to conduct inference on government revenue from ad valorem taxes ν across markets with the map $F(h(\theta))=\sum_{j\in\mathcal{J}}\nu_j h_j(\theta)s_j(h(\theta))$. We have:*

$$\left\{G(h(\theta_0))\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)'G(h(\theta_0))'/N\right\}^{-1/2}(F(h(\hat{\theta}_N))-F(h(\theta_0)))\xrightarrow{d}N(0,I),$$

where $G(h(\theta))=(\nu\odot s(h(\theta))'+(\nu\odot h(\theta))'D(h(\theta))$ where $D(\cdot)$ is stacked block-diagonal by markets.

Proof. Taking the gradient of the map F with respect to h, we find the j-th element as:

$$\begin{aligned}[\nabla_h F(h(\theta))]_j&=\nu_j[s_j(h(\theta))+h_j(\theta)[\nabla_h s(h(\theta))]_{jj}]+\sum_{k\neq j}\nu_k h_k(\theta)[\nabla_h s_k(h(\theta))]_{kj}\\&=\nu_j s_j(h(\theta))+\sum_k\nu_k h_k(\theta)[D(h(\theta))]_{kj}.\end{aligned}$$

Notice that, by market independence, gradients with respect to h across markets are zero. Stacking the gradients $D(\cdot)$ block-diagonal across markets, we have:

$$G(h(\theta))\equiv\nabla_h F(h(\theta))=(\nu\odot s(h(\theta))'+(\nu\odot h(\theta))'D(h(\theta)).$$

The result follows immediately from Theorem 8.

□

Corollary 5. *Denote $\mathcal{J}=\mathcal{J}_1\cup\cdots\cup\mathcal{J}_M$ as the union of product sets across markets $m=1,...,M$. Suppose we would like to conduct inference on government revenue from unit taxes τ across markets with the map $F(h(\theta))=\sum_{j\in\mathcal{J}}\tau_j s_j(h(\theta))$. We have:*

$$\left\{\tau' D(h(\theta_0))\nabla_\theta h(\theta_0)\Omega_0^{-1}\nabla_\theta h(\theta_0)'D(h(\theta))'\tau/N\right\}^{-1/2}(F(h(\hat{\theta}_N))-F(h(\theta_0)))\xrightarrow{d}N(0,I),$$

where $\nabla_h F(h(\theta)) = \tau' D(h(\theta))$ *and* $D(\cdot)$ *is stacked block-diagonal across markets.*

Proof. Taking the gradient of the map F with respect to h, we find the j-th element as:

$$\begin{aligned}[\nabla_h F(h(\theta))]_j &= \tau_j [\nabla_h s(h(\theta))]_{jj}] + \sum_{k \neq j} \tau_k [\nabla_h s_k(h(\theta))]_{kj} \\ &= \sum_k \tau_k [D(h(\theta))]_{kj}.\end{aligned}$$

Notice that, by market independence, gradients with respect to h across markets are zero. Stacking the gradients $D(\cdot)$ block-diagonal across markets, we have $\nabla_h F(h(\theta)) = \tau' D(h(\theta))$. The result follows immediately from Theorem 8.

□

F Simulation Details and Additional Results

We discuss additional details concerning our simulation environments and computational details used in implementing our Monte Carlo experiments. We proceed in five steps. First, we characterize the differences between the three simulation environments both in terms of market structure and demand. Second, we discuss the supply-side models used to generate data in all three environments. Third, we provide details on estimating VMM and the parametric models of supply. Fourth, we explain how we compute counterfactuals under the estimated VMM and parametric model. Fifth, we discuss the steps to quantify uncertainty arising in the predicted counterfactuals. Table F1 shows which environments and exercises are useful for each step.

F.1 Details on Constructing Simulation Environments

Baseline Environment: We generate samples with $T \in \{100, 1000, 10000\}$ markets. In each market, we start with three products and then randomly drop at most one product from each environment. As we only consider single product firms, this results in approximately half the markets in a dataset being duopoly markets and half being tripoloy markets. For demand, we adopt a logit specification. Consumer i derives utility from product j in market t according to:

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

where x_{jt} represents a constant and two observed product characteristics $x_{jt}^{(1)}$ and $x_{jt}^{(2)}$. ξ_{jt} captures unobserved quality, and ϵ_{ijt} follows a Type I extreme value distribution. $x_{jt}^{(1)}$ and $x_{jt}^{(2)}$ are independently drawn from a $U[0,1]$ distribution. ξ_{jt} is drawn from a $U[0,1]$ distribution and has correlation $\rho = 0.9$ with unobserved supply shocks ω_{jt}, which is also drawn from $U[0,1]$. For Laffer

TABLE F1: Simulation Design Summary

Exercise	Env.	T	Conduct	Cost
Predictive Accuracy in Hold-Out Samples				
Holdout Performance (Sec. 5.2)	Base	100/1K/10K	B, PW	C, E
Scalability				
High-Dimensional (Sec. 5.3)	HiDim	100/1K	B, PW	C
Economic Interpretability				
Pass-Through (Sec. 5.4)	Base	1K	B, PW	C
Policy Counterfactuals				
Characteristic Regulation (Sec. 5.5)	Base	1K	B, PW	C, E
Tax Policy (Sec. 5.5)	Base	1K	B, PW	C
Merger Simulation (Sec. 5.5)	Merger	1K	B, PW	C
Uncertainty Quantification				
Inference (Sec. 5.6)	Base	100/1K	B, PW	C

Notes: Env. = Environment: Base = Baseline (1-3 products); HiDim = High-dimensional (30 products); Merger = Merger-specific (10-15 products, multi-product firms). Conduct: B = Bertrand-Nash; PW = Profit-weight. Cost: C = Constant marginal cost; E = Economies of scale. Multiple values separated by commas indicate separate results for each specification.

curve counterfactuals, we augment the environment with variation in taxes. Crucially for identification of the flexible supply function, markets feature variation in both ad valorem and unit taxes. Ad valorem taxes are drawn from $\nu_t \sim U[0, 0.6]$ while unit taxes follow $\tau_{jt} \sim U[4, 8]$.

High-Dimensional Environment: To address concerns about the curse of dimensionality inherent in nonparametric methods, we implement a second environment inspired by the empirical setting in Miller and Weinberg (2017). Following the market structure of the US beer market in that paper, this environment features 30 differentiated products offered across markets, an order of magnitude larger than our basic setup and representative of product variety in many IO settings. To generate data, we start with five firms in every market, each producing six products. We then randomly drop up to ten products, so that the final datasets contain 20-30 products in each market. For demand, we adopt a nested logit demand system, which is more flexible than in our basic environment. Specifically, we adopt an inside-outside nesting structure so that the utility individual i receives from inside product j in market t is given by:

$$u_{ijt} = \alpha p_{jt} + \beta x_{jt} + \xi_{jt} + \zeta_{it} + (1 - \sigma)\epsilon_{ijt},$$

where x_{jt} represents a constant and a single observed product characteristic $x^{(1)}$. ξ_{jt} captures unobserved quality, ζ_{it} captures the random taste for inside products following the Cardell distribution Cardell (1997), and ϵ_{ijt} follows a Type I extreme value distribution. To best match the environment

in Miller and Weinberg (2017) we treat $x_{jt}^{(1)}$ like their month-product fixed effect, setting $\beta = 1$ and drawing $x_{jt}^{(1)}$ from $N(0, 0.2)$ which approximates the empirical distribution of their estimated month-product fixed effects. We also follow Miller and Weinberg (2017) in setting $\alpha = -0.0887$ and $\sigma = 0.83$. The unobserved demand anf cost shocks ξ_{jt} and ω_{jt} are jointly drawn from a $U[0, 1]$ distribution with variance-covariance matrix:

$$V = \begin{bmatrix} 0.18 & 0.04 \\ 0.04 & 1.08 \end{bmatrix},$$

which matches the empirical variance-covariance matrix in Miller and Weinberg (2017).

Merger Simulation Environment: For merger simulation counterfactuals (see Section 5.5), we augment the high-dimensional environment to create richer variation in market structure and ownership matrix $\mathcal{H}_t$. In doing so, the pre-merger data contains variation analogous to the merger we simulate, helping the flexible model learn how the proposed change in market structure will affect equilibrium pricing. In 50% of the markets we have three firms, one with 6 products, one with 5 products and one with 4 products. In the other 50% of markets, we have 4 firms, one with 5 products, one with 4 products and 2 with 3 products. We randomly drop up to 5 products from each market in the same manner as in the high-dimensional environment. Here, we utilize the exact demand system in the baseline environment.

F.2 Details on Supply-Side Models in the DGPs

Marginal Cost Specifications: In all three data-generating environments, we generically parameterize the marginal cost as

$$c_{jt} = w_{jt}'\gamma + \lambda_0 s_{jt} + \lambda_1 s_{jt}^2 + \omega_{jt}$$

In all simulations in the baseline and merger-simulation environments, w_{jt} contains a constant and two observed cost shifters, excluded from x_{jt}, which are drawn iid from a $U[0, 1]$ distribution. As discussed above, the unobserved cost shock ω_{jt} is drawn jointly with the unobserved demand shock ξ_{jt} from a standard bivariate uniform distribution with correlation coefficient $\rho = 0.9$ (the default in `pyblp`). In these environments, we set the parameters $\gamma = [3, 6, 4]$. The values of λ_0 and λ_1 allow us to control whether the supply model exhibits economies of scale. In DGPs imposing constant marginal cost, $\lambda_0, \lambda_1 = 0$; for economies of scale specifications, $\lambda_0 = 20$ and $\lambda_1 = -30$.

In simulations under the high-dimensional environment, we instead draw a single observed cost shifter from an exponential distribution with scale parameter set to 1.25 to best match the empirical distribution of the observed cost shifter in Miller and Weinberg (2017). For $w_{jt} = (1, w_{jt}^{(1)})$ we set $\gamma = [6.809, 0.168]$. ω_{jt} is drawn jointly with ξ_{jt} according to the uniform distribution described

above so that the variance-covariance matrix approximates that in Miller and Weinberg (2017). In the high-dimensional environment, we only consider constant marginal costs so that $\lambda_0, \lambda_1 = 0$.

Supply-side Models: For the supply side, the generic first-order conditions generating prices can be expressed as:

$$p_{jt} = (\mathcal{H}_t \odot D_t)^{-1} s_t + c_{jt}.$$

When the supply-side model generating the data involves Nash-Bertrand conduct, the (j,k)-th element of the ownership matrix $\mathcal{H}_t$ is zero when products j and k are produced by different firms in market t and one otherwise. For profit-weight models, the (j,k)-th element of $\mathcal{H}_t$ is parameterized as κ when products j and k are produced by different firms in market t. In the baseline environment, $\kappa = 0.5$; in the High-Dimensional and Merger-Sim Environments, we set $\kappa = 0.75$.

Supply-side Models for Laffer Curve Exercises: As discussed above, we modify the baseline environment when generating data to perform the Laffer curve counterfactual by incorporating market-level ad valorem taxes (v_t) and unit taxes (τ_{jt}) into our simulations. We impose unit taxes directly on firms, while ad valorem taxes are levied on consumers. Defining p_{jt} as the tax-inclusive price paid by consumers, the fraction of the consumer's payment received by the firm is $\nu_t = 1/(1+v_t)$. The augmented after-tax first-order conditions generating firms' prices are given by:

$$\nu_t p_{jt} - \tau_t = \nu_t \Delta_{0jt} + c_{jt}, \tag{F1}$$

where $\nu_t p_{jt} - \tau_t$ is the after-tax revenue received by the firm for product j in market t.

F.3 Estimation Details

Variational Method of Moments Deep neural networks require the specification of a number of hyperparameters. Our implementation of VMM is no exception: we use two deep neural networks to fit our flexible supply model. We enumerate the fine-tuning that we used for the paper. For many of the hyperparameters, we leave the defaults from Bennett and Kallus (2023). We use the Optimistic Adam (OAdam) algorithm to train the deep neural networks, setting the learning rate to $\eta = 5 \times 10^{-4}$ and decay of momentums to $\beta = [0.5, 0.9]$. We do not perform gradient descent by batching markets, instead opting to include all markets in a single batch.[25] During training, we track loss for a smaller hold-out validation sample. Throughout, we implement early stopping by saving the weights of the deep neural networks that correspond to the lowest validation loss.[26] We do not include a regularization term R, leaving any regularization to our early early stopping procedure. While we experiment with the dimensionality of the primary deep neural network h, we fix the dimensionality of the deep neural network f to that of Bennett and Kallus (2023): two

[25] We found that this improved speed while leaving performance relatively unaffected.

[26] We additionally define ex ante early stopping rule but find that performance is comparable.

hidden layers, the first with 50 nodes and the second with 20 nodes. Both of our deep neural networks f, h are fully connected with rectified linear units (ReLU) activation functions.

In the training process, we require specifications of endogenous, exogenous, and instrumental variables. The endogenous variables include the vector of J_t market shares and the matrix of $\frac{J_t(J_t-1)}{2} + J_t$ demand derivatives for each market t. The exogenous variables include K^w own cost shifters w_{jt}, where K^w is the number of observable cost shifters. For instruments, we include: (i) $K^x \times J_t$ own and rival product characteristics, where K^x is the number of product characteristics that enter demand and are excluded from cost; (ii) $K^w \times (J_t - 1)$ rival cost shifters; (iii) and $1 + K^x + K^w$ constructed BLP instruments for other products and rival firms. In conjunction with the manifold completeness assumption for identification, we maintain sufficient dimensionality of instruments for identification.

Parametric Models For each parametric model, we estimate the model by imposing a model of conduct to invert marginal costs and then project marginal costs on observable cost shifters. We walk through each of the models in turn, beginning with the simplest specifications. Under perfect competition, we assume that firms price at marginal cost, meaning prices are equal to marginal costs. We regress these costs on observable cost shifters to recover the implied parameters and residuals. Next, turning to price-setting models (Bertrand, profit-weight, and monopoly), we assume a model of conduct with its corresponding ownership matrix and use the first-order conditions to invert marginal costs (described above). Again, we regress these costs on observable cost shifters to recover implied parameters and residuals. In some specifications of the data-generating process, we include economies of scale. Under economies of scale, we include market shares and squared market shares in the regression. To address the endogeneity present in the problem, we instrument these two variables with own product characteristics and squared own product characteristics under the assumption that they are excluded from the marginal cost. We recover the implied parameters and residuals, as before.

F.4 Hold-out Sample, Cost Pass-Through, and Counterfactual Performance

We now turn to our evaluation of the parametric and flexible models. Before doing so, it is useful to establish the computation of equilibrium prices. For parametric model m, we solve the fixed point in prices using Equation (F2) below:

$$\tilde{\nu}_t \tilde{p}_{jt} - \tilde{\tau}_t = \tilde{\nu}_t \tilde{\Delta}_j^m(\tilde{p}_t, \tilde{D}(\tilde{p}_t, \tilde{x}_t, \tilde{\xi}_t), \tilde{\mathscr{H}}_t) + \tilde{c}_j^m(\tilde{q}_t, \tilde{\mathrm{w}}_{jt}, \tilde{\omega}_{jt}), \qquad j = 1, ..., J, \tag{F2}$$

where variables with tildes can be altered in the counterfactual. We use the fixed point procedure in Morrow and Skerlos (2011) to solve for prices. Solving for equilibrium prices in the flexible model

is similar. We solve a modified fixed point in prices described by Equation (F3):

$$\nu_t \tilde{p}_{jt} - \tilde{\tau}_t = \hat{h}(\jmath(\tilde{p}_t, \tilde{x}_t, \tilde{\xi}_t), D(\tilde{p}_t, \tilde{x}_t, \tilde{\xi}_t), \tilde{\mathrm{w}}_{jt}, \tilde{\nu}_t, \tilde{\mathscr{H}}_t) + \tilde{\hat{\omega}}_{jt}, \qquad j = 1, ..., J, \tag{F3}$$

where $\hat{\omega}_{jt}$ are the residuals recovered for j, t from the estimated function $\hat{h}$. $\tilde{\hat{\omega}}_{jt}$ are potentially altered residuals from the flexible supply function. To solve the fixed point, we use a root-finding technique. We utilize `df-sane` in the `scipy` package, a derivative-free spectral method with a tolerance of 1×10^{-6}. Unless otherwise noted below, all variables keep their values without tildes.

Hold-Out Performance After obtaining estimates of our flexible supply model and the parametric supply models, we first compare their performance at predicting prices in a hold-out sample. For each parametric model m, we have estimates $\hat{\gamma}^m$ of the parameters in marginal cost. For observations in the hold-out sample, we recover the implied marginal costs c_{mjt} by invert the model-implied first order conditions. Then, we recover the model-implied unobservable cost shocks from the implied marginal costs as $\hat{\omega}_{mjt} = c_{mjt} - w'_{jt}\hat{\gamma}^m$. For the flexible model of supply, we compute the model-implied unobservable cost shifter by predicting prices from the fitted model and comparing them to observed prices in the hold-out sample. We compute the mean-squared error (MSE) for each of the models; the closer the MSE is to the irreducible error in the true data-generating process (the MSE of the true model), the better the performance.

Computing Cost Pass-through In the baseline environment with constant marginal costs, we also compare the cost pass-through of our estimated flexible supply model to the cost pass-through implied by the true model. In each market, we compute a numerical approximation to the cost pass-through matrices under a given model. Specifically, we obtain the columns of each pass-through matrix by iterating over the products in the market.[27] For each product, we increase its model-implied marginal cost (which corresponds to the true marginal cost under the true model) by 1 percent. We then recompute the equilibrium price vector across all firms in that market. For the true parametric models, we use the algorithm developed in Morrow and Skerlos (2011) to solve the fixed point in Equation F2 to obtain these counterfactual price vectors. For our estimated flexible model, we perturb the estimated residual $\hat{\omega}_{jt}$ by adding 1 percent of the true marginal cost to $\hat{\omega}_{jt}$, denoted $\tilde{\hat{\omega}}_{jt}$. Under the perturbed $\tilde{\hat{\omega}}_{jt}$, we solve for the vector of equilibrium prices in market t by solving a fixed point described by Equation F3. Under the parametric and flexible supply models, we report the change in prices divided by the change in cost.

We now explain how we modify markets when computing market counterfactual outcomes. For each counterfactual, we solve fixed points in prices under a specified parametric model m using Equation F2 and under the flexible model of supply using Equation F3. Unless otherwise noted,

[27] The (j, k)-th element of each pass-through matrix corresponds to the change in price of product j associated with a marginal change in the cost of product k.

all variables remain the same as they did during estimation.

Characteristic Regulation In the first set of market counterfactuals, we evaluate regulations on characteristics that enter the model as cost shifters in the baseline environment. For the parametric model m, we recover costs and perform a regression on observed cost shifters to recover coefficients $\hat{\gamma}^m$ with corresponding residuals $\hat{\omega}^m$. In the case of economies of scale, we additionally regress on observed market shares s_{jt} and squared market shares s_{jt}^2, instrumenting with product characteristics x_{jt} and squared product characteristics x_{jt}^2 to address endogeneity. For the flexible model, we use the estimated supply function $\hat{h}$ and its corresponding residual.

For this set of counterfactuals, we alter the first indexed cost shifter $\tilde{\mathrm{w}}_{jt}^1 = \mathrm{w}_{jt}^1 + 1$, recalling that w_{jt}^1 is drawn iid $U[0,1]$, meaning it is pushed well outside the support in the training data. To evaluate parametric market counterfactuals, we use the estimated $\hat{\gamma}^m$ and $\hat{\omega}^m$ to predict costs with the modified $\tilde{\mathrm{w}}_{jt}$:

$$\tilde{c}_{jt}^m = \hat{\gamma}^m \tilde{\mathrm{w}}_{jt} + \hat{\omega}_{jt}^m,$$

and solve for equilibrium prices using Equation F2. Similarly, for the flexible supply model, we input altered $\tilde{\mathrm{w}}_{jt}$ into the function $\hat{h}$ and use the fixed point in Equation F3 to solve for equilibrium prices. In both parametric and flexible supply models, market shares and demand derivatives are updated at each iteration of the fixed point to accurately reflect the state of demand at the guess of prices. Under economies of scale, costs can shift with market shares as well. We report differences in prices for this set of counterfactuals.

Product Characteristic Regulation The next set of counterfactuals is similar to the first, but requires a different procedure to solve the parametric equilibrium. Again, we perform these counterfactuals in the baseline environment. We alter the first indexed product characteristic $\tilde{x}_{jt}^1 = x_{jt}^1 + 1$, recalling that x_{jt}^1 is drawn iid $U[0,1]$, meaning it is again pushed well outside the support in the training data. We then plug in the altered $\tilde{x}_t$ and solve for equilibrium prices using Equation F2 and Equation F3, fixing cost parameters and shocks as above. We report differences in shares for this set of counterfactuals, using the demand system evaluated at the set of equilibrium prices.

Product Exit Product exit counterfactuals are implemented for the baseline environment. We implement exit by dropping the first firm from each market in which it is present, altering the ownership matrix to $\tilde{\mathcal{H}}$ accordingly. Under the new ownership matrix, we solve for equilibrium prices using Equation F2 and Equation F3, fixing cost parameters and shocks as above. In the flexible model, we pad the data with zeros to maintain the correct input dimension. Variation in ownership structure is thus very helpful for the deep neural networks to learn the supply function, especially with high-dimensional product spaces. We report differences in consumer surplus across all products and markets using compensating variation as defined in Small and Rosen (1981).

Laffer Curve We implement Laffer curve counterfactuals in the baseline environment with constant marginal cost for two sets of tax instruments: unit and ad valorem taxes. These instruments are implemented in isolation; that is, we only consider an environment in which only one is implemented, not both. Taxes are assumed to be the same across products. For training, we introduce variation in taxes across markets. Under the unit tax $\tilde{\tau}_t$, we solve for parametric equilibrium prices in Equation F2 by adding $\tilde{\tau}_t$ to the marginal cost. For the flexible model, we load the tax $\tilde{\tau}_t$ onto the residuals such that $\tilde{\hat{\omega}}_{jt} = \hat{\omega}_{jt} + \tilde{\tau}_t$ and solve for equilibrium prices using Equation F3. Laffer curves are constructed by computing government revenue in a representative market t:

$$G_t^U = \sum_{j \in J_t} \tilde{\tau}_t \tilde{s}_{jt},$$

where $\tilde{s}_{jt}$ are the resulting market shares from the equilibrium prices under unit tax $\tilde{\tau}_t$. Laffer curves are then produced by mapping government revenue G_t^U over the dollar amount of the unit tax τ_t.

For ad valorem taxes, we solve the parametric models by setting counterfactual marginal costs according to the tax rate:

$$\tilde{c}_{jt}^m = \frac{1}{\tilde{\nu}_t} \tilde{c}_j^m(\mathrm{w}_{jt}, \omega_{jt}^m),$$

where $\tilde{\nu}_t = 1/(1+\tilde{v}_t)$ and $\tilde{v}_t$ is the ad valorem tax. For the flexible supply model, we incorporate the tax rate v_t directly as an exogenous variable in training and thus update this value to $\tilde{v}_t$. We then use Equation F3 to solve for equilibrium prices under the counterfactual tax rate. Laffer curves are constructed again by computing government revenue in a representative market t:

$$G_t^A = \sum_{j \in J_t} (1 - \nu_t) \tilde{p}_{jt} \tilde{s}_{jt},$$

where $\tilde{s}_{jt}$ are the resulting market shares from the equilibrium prices under ad valorem tax $\tilde{v}_t$. Laffer curves are then produced by mapping government revenue G_t^A over the tax rate of the ad valorem tax ν_t.

Merger Simulation Merger simulation counterfactuals are implemented for the merger simulation environment. We implement merger simulation by unifying the firm identifiers for the two merging firms (i.e., the two three-product firms) and altering the ownership matrix $\tilde{\mathcal{H}}_t$ accordingly. Under the new ownership matrix, we solve for equilibrium prices using Equation F2 and Equation F3, fixing cost parameters and shocks as above. In the flexible model, we pad the data with zeros to maintain the correct input dimension. Variation in ownership structure is thus very helpful for the deep neural networks to learn the supply function, especially in this high-dimensional environment. It is important to notice that, by construction, the ownership structure in markets affected by the merger is "in-sample" in the sense that this exact market structure was observable during training. We report differences in consumer surplus across all products and markets using compensating

variation as defined in Small and Rosen (1981).

F.5 Quantifying Uncertainty

Quantification of uncertainty also requires the specification of hyperparameters because it involves training another deep neural network. As in estimation, we utilize OAdam as our choice of optimization algorithm, setting the learning rate to $\eta = 5 \times 10^{-2}$ and decay of momentums to $\beta = [0.5, 0.9]$. We include all markets in a single batch during gradient descent and implement a smooth start by averaging over the last 4,000 epochs of training. The dimensionality of f mirrors that of estimation with 50 nodes in the first hidden layer and 20 nodes in the second. We omit regularization terms. Again, f is fully connected with leaky ReLU activation functions. In our computation of $\nabla_\theta \omega(\theta_0)$, we utilize automatic differentiation native to `torch` to differentiate the loss function with respect to the parameters of the deep neural network, stacking them into a $b \times 1$ vector consistent with the variance objective function. In Algorithm 1, we take T_α from a folded normal distribution with tuning parameter $c = 1$ and $\alpha = 0.05$ as our fixed critical values. In the figures throughout, we adapt the Bonferroni correction as a more conservative approach for the sake of interpretability; our inference algorithm generates strictly tighter confidence intervals.

TABLE G1: MSE Across Models, Single-Product Environment with Constant Costs

T	True Model	Standard Models			Flexible Models			D_t included
		B	M	P	$h=3$	$h=20$	$h=100$	
Panel A: Bertrand DGP								
100	0.95	0.89	1,688.01	4.63	1.59	1.53	1.03	No
					1.34	1.01	1.12	Yes
1,000	0.98	0.96	1,589.55	10.00	2.37	1.03	1.11	No
					1.22	1.19	1.17	Yes
10,000	0.99	0.99	3,001.09	9.81	2.76	1.07	1.01	No
					1.23	0.99	0.97	Yes
Panel B: Profit-Weight DGP								
100	0.95	3.65	85.86	6.33	1.71	1.23	1.65	No
					1.17	1.29	1.34	Yes
1,000	0.98	4.40	65.32	9.90	2.72	1.49	1.16	No
					1.27	0.94	0.96	Yes
10,000	0.99	4.53	118.37	9.88	3.02	2.29	1.06	No
					1.26	1.01	0.99	Yes

Notes: The table reports the mean squared error (MSE) in prices for the true model, a set of standard parametric models (B = Bertrand with constant cost, M = joint profit maximization with constant cost, P = marginal cost pricing with constant cost), and the flexible model. In Panel A, the true supply model generating the data is Bertrand with constant marginal cost, while in Panel B, it is a profit-weight model with constant marginal cost. Each DGP is from the single-product environment. The Flexible Model columns include neural networks of varying dimensions. For example, $h = 3$ refers to a 3×3 hidden layer. For each imposed supply model under each DGP, the MSE in predicted prices is computed on a hold-out test sample.

G Additional Simulation and Counterfactual Results

G.1 Additional Hold-Out Sample Results

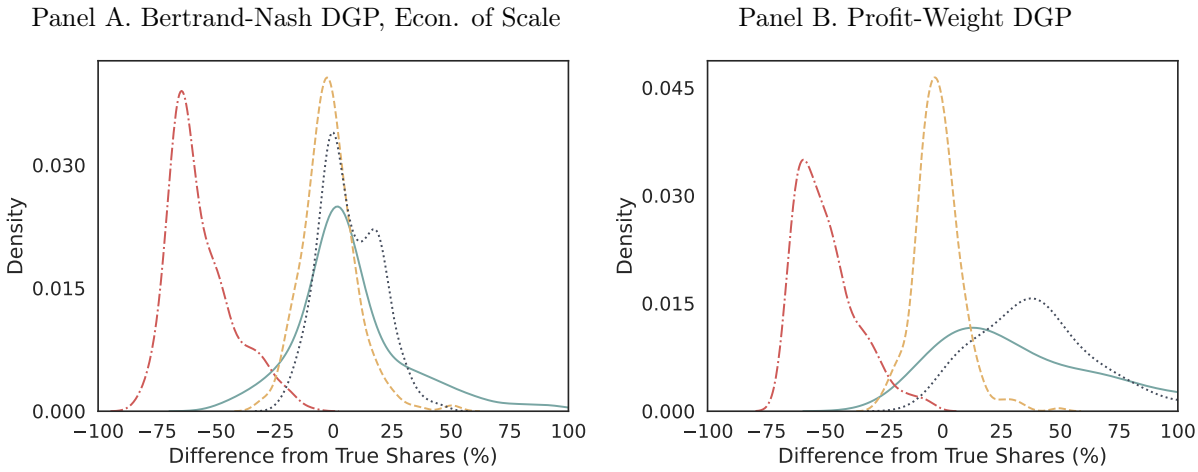
TABLE G2: MSE Ranges Across Models

T	Network Size	D_t Included	P5	P25	P50	P75	P90
Panel A: Bertrand DGP							
100	Small (h=3)	No	1.40	1.50	1.54	1.56	1.58
		Yes	1.02	1.12	1.32	1.53	1.72
	Medium (h=20)	No	1.09	1.20	1.48	1.54	1.57
		Yes	1.00	1.02	1.04	1.06	1.10
	Large (h=100)	No	0.99	1.10	1.18	1.27	1.56
		Yes	1.00	1.09	1.15	1.21	1.35
1,000	Small (h=3)	No	2.29	2.35	2.37	2.40	2.43
		Yes	1.18	1.20	1.21	1.23	2.00
	Medium (h=20)	No	0.99	1.01	1.02	1.03	2.29
		Yes	0.99	1.01	1.11	1.19	1.21
	Large (h=100)	No	1.00	1.05	1.10	1.14	1.23
		Yes	0.98	1.00	1.03	1.07	1.15
Panel B: Profit-Weight DGP							
100	Small (h=3)	No	1.49	1.55	1.59	1.65	1.71
		Yes	1.19	1.28	1.36	1.46	2.10
	Medium (h=20)	No	1.14	1.30	1.50	1.60	1.71
		Yes	0.88	1.24	1.35	1.45	1.52
	Large (h=100)	No	1.22	1.35	1.42	1.56	1.71
		Yes	1.09	1.27	1.39	1.51	1.69
1,000	Small (h=3)	No	2.71	2.73	2.75	2.77	2.80
		Yes	1.23	1.24	1.26	1.32	2.04
	Medium (h=20)	No	1.06	1.12	1.27	1.49	1.61
		Yes	0.94	0.97	0.98	0.99	1.01
	Large (h=100)	No	1.00	1.03	1.07	1.10	1.17
		Yes	0.94	0.97	0.99	1.01	1.04

Notes: The table reports the moments of distribution of mean squared errors (MSE) in prices for the VMM model with 100 different neural network weight initializations. The models are trained on duopolies and triopolies. The MSE is computed on a hold-out test sample restricted to markets where merging firms are present. Panel A reports the results for a Bertrand DGP for different sizes of neural networks. Panel B likewise reports results for the profit-weight DGP ($\kappa = 0.5$).

G.2 Additional Counterfactual Results

FIGURE G1: Regulation of Product Characteristics: Share Predictions

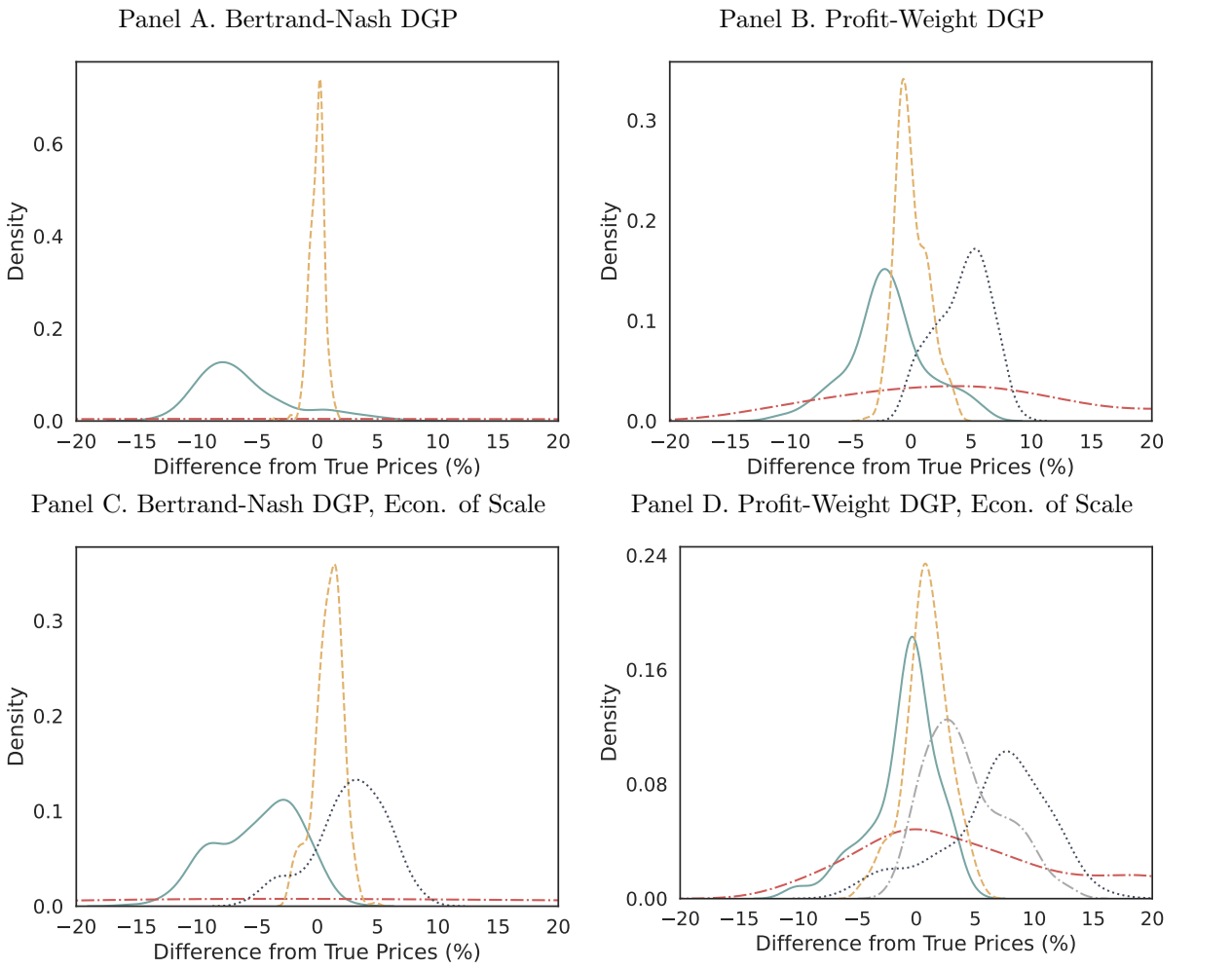


Fitted Model — DGP

	A. Bertrand	B. Profit-Weight	C. Bertrand (Scale)	D. Profit-Weight (Scale)
Bertrand (Scale)	-	-	-	77.3
Bertrand (Const.)	-	53.51	14.27	100.58
Monopoly	57.88	51.22	58.26	44.08
Perf. Comp.	23.29	59.72	24.94	77.59
Flex Supply	3.65	10.07	11.59	17.57

Notes: The figure displays share prediction errors when characteristic $\tilde{x}_1 = x_1 + 1$ is pushed out of its training support. The flexible model uses a medium neural network with demand derivatives on $T = 1,000$ markets.

FIGURE G2: Regulation of Cost Shifters: Price Predictions

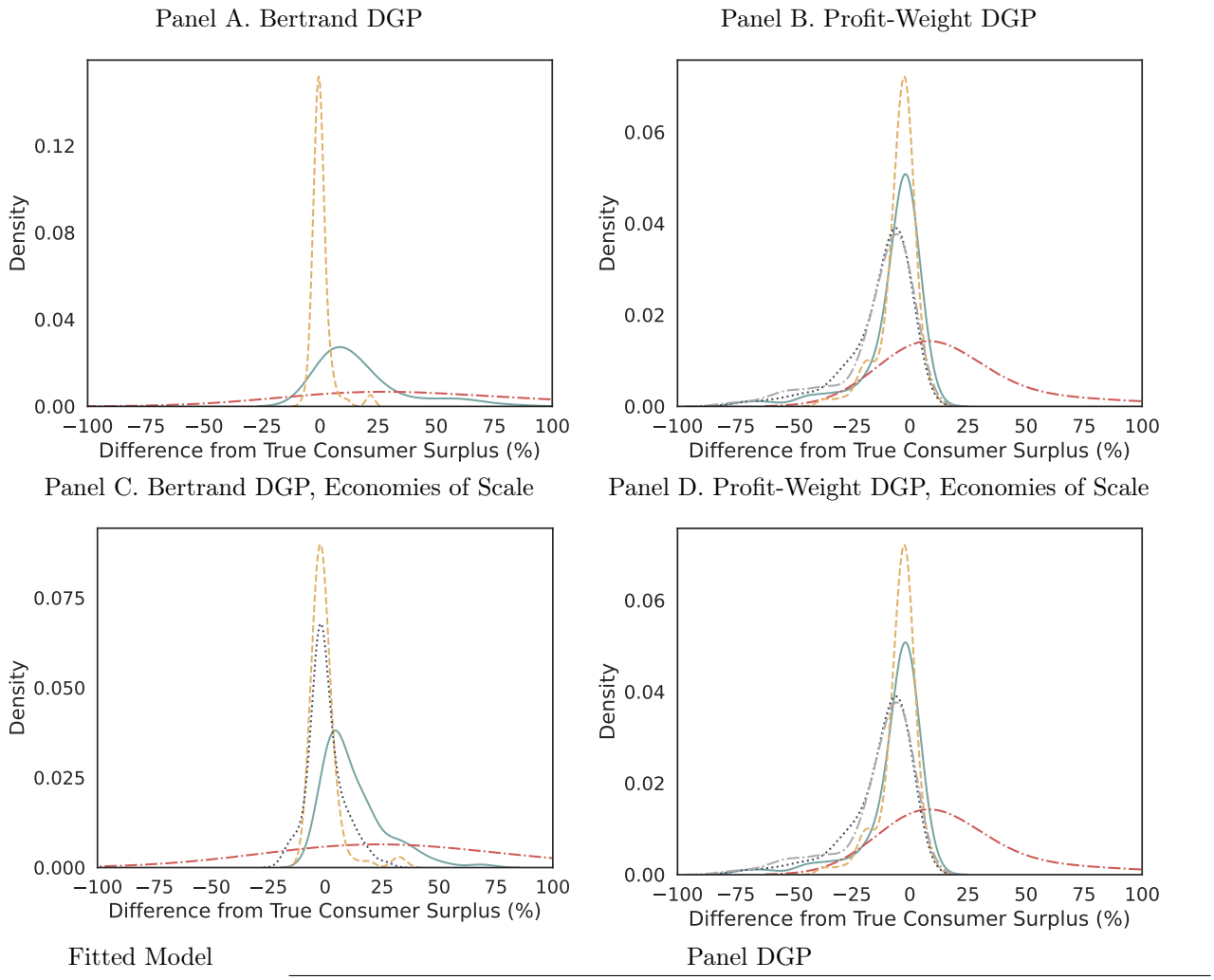


Fitted Model | Panel DGP

	A. Bertrand	B. Profit-Weight	C. Bertrand (Scale)	D. Profit-Weight (Scale)
Bertrand (Scale)	-	-	-	5.28
Bertrand (Const.)	-	4.82	4.08	8.08
Monopoly	105.47	12.96	59.41	10.20
Perf. Comp.	7.25	3.81	6.01	3.11
Flex Supply	0.62	1.34	1.50	2.10

Notes: The figure displays price prediction errors when characteristic w_1 is doubled. The flexible model uses a medium neural network with demand derivatives on $T = 1,000$ markets.

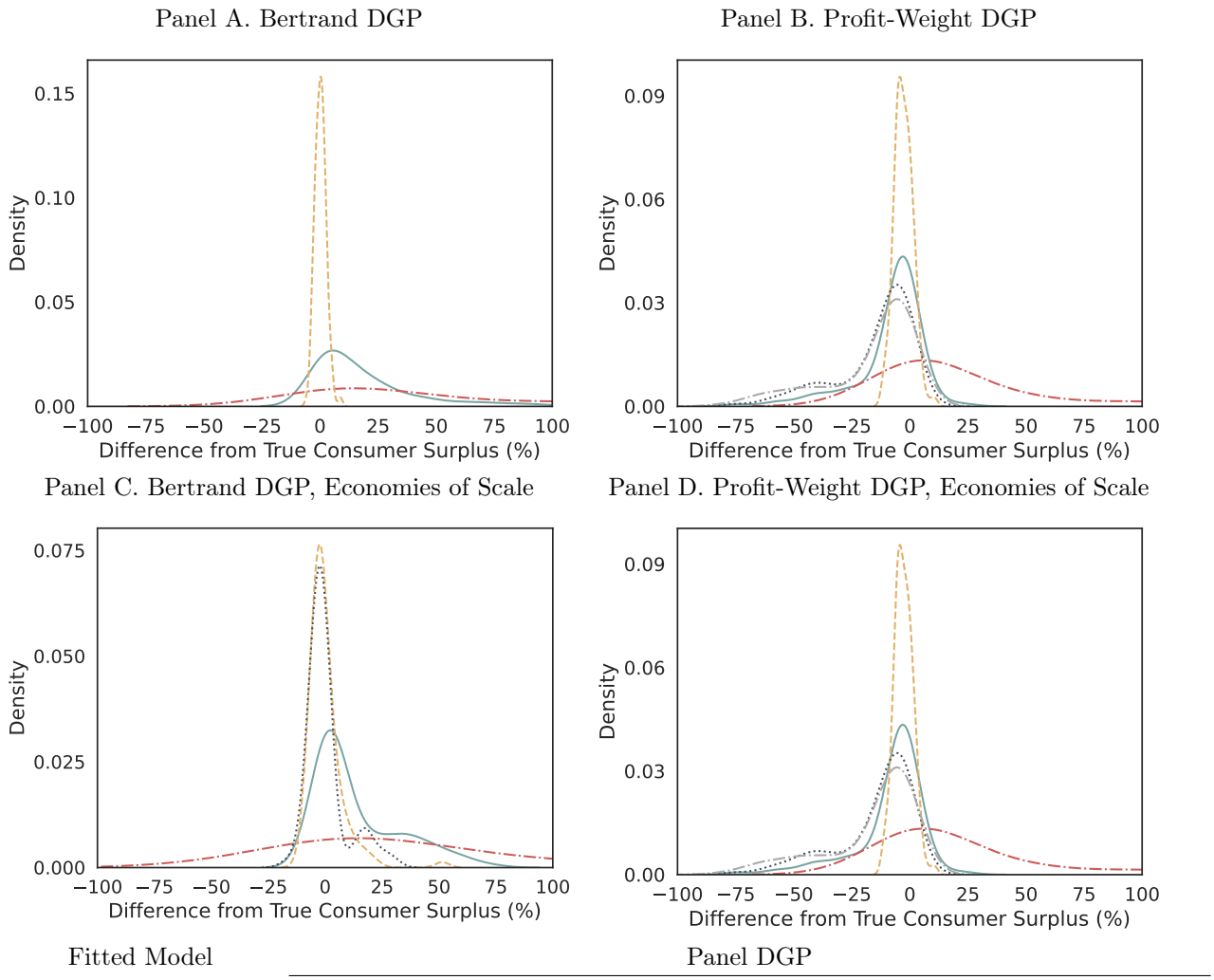
FIGURE G3: Single-Product Merger Simulation



	A. Bertrand	B. Profit-Weight	C. Bertrand (Scale)	D. Profit-Weight (Scale)
Bertrand (Scale)	-	-	-	21.10
Bertrand (Const.)	-	22.70	7.69	19.67
Monopoly	113.77	44.28	130.82	57.57
Perf. Comp.	28.51	20.25	18.55	15.51
Flex Supply	4.24	2.64	6.66	9.06

Notes: The figure and table show the distribution and mean squared error (MSE) in prices for the true model (Bertrand and profit-weight with $\kappa = 0.5$), standard models, and the flexible supply model in the single-product merger simulation exercise. Panels A and B correspond to results for Bertrand and profit-weight DGPs without economies of scale, while Panels C and D include economies of scale. The model is trained on duopolies and triopolies. The MSE is computed on a hold-out test sample restricted to markets where merging firms exist. The neural network used in estimation is small with a 3×3 hidden layer. The models are trained on $T = 1,000$ markets. Profit-weight models include demand derivatives.

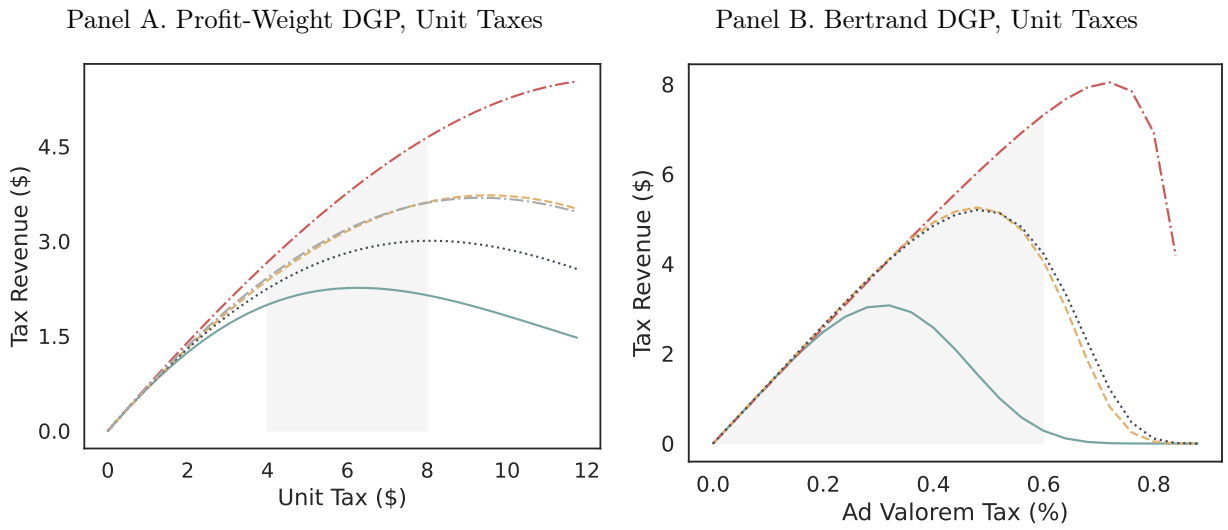
FIGURE G4: Single-Product Merger Simulation (Triopolies)



	A. Bertrand	B. Profit-Weight	C. Bertrand (Scale)	D. Profit-Weight (Scale)
Bertrand (Scale)	-	-	-	25.71
Bertrand (Const.)	-	23.21	8.65	22.42
Monopoly	88.92	40.54	119.37	61.51
Perf. Comp.	27.72	19.62	23.08	17.39
Flex Supply	2.37	2.19	7.77	4.69

Notes: The figure and table show the distribution and mean squared error (MSE) in prices for the true model (Bertrand and profit-weight with $\kappa = 0.5$), standard models, and the flexible supply model in the single-product merger simulation exercise. Panels A and B correspond to results for Bertrand and profit-weight DGPs without economies of scale, while Panels C and D include economies of scale. The model is trained on triopolies. The MSE is computed on a hold-out test sample restricted to markets where merging firms exist. The neural network used in estimation is small with a 3×3 hidden layer. The models are trained on $T = 1,000$ markets. Profit-weight models include demand derivatives.

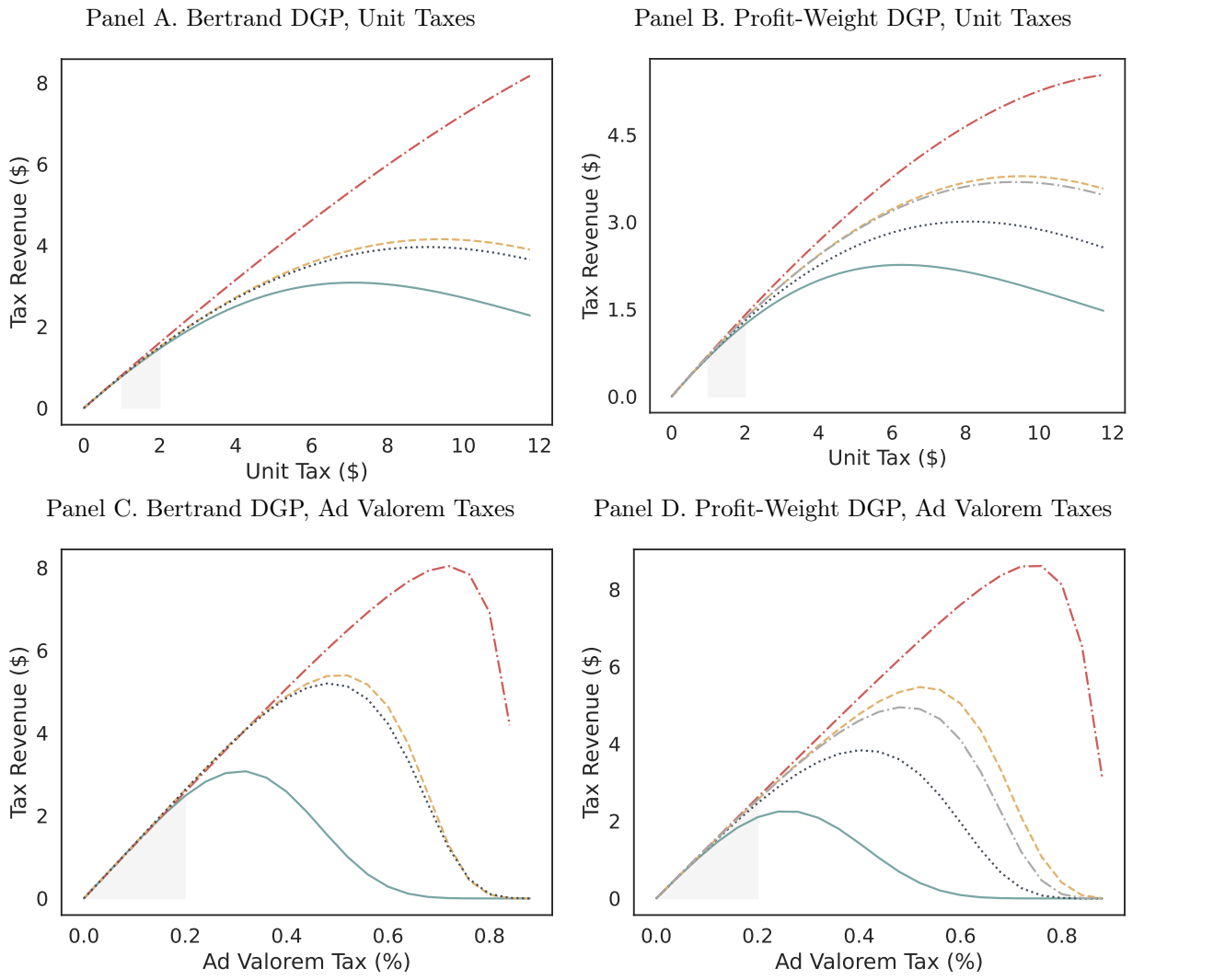
FIGURE G5: Laffer Curves for Unit and Ad Valorem Taxes



Fitted Model	DGP			
	A. Bertrand (U)	B. Profit-Weight (U)	C. Bertrand (AV)	D. Profit-Weight (AV)
Bertrand	-	0.51	-	1.02
Monopoly	2.07	0.99	3.28	3.82
Perf. Comp.	0.75	1.19	2.12	2.41
Flex Supply	0.04	0.04	0.15	0.02

Notes: The figure displays Laffer curves for the flexible model estimated with VMM and a set of standard models under different conduct assumptions. Unit taxes are drawn from the uniform distribution $U[4, 8]$ and ad valorem tax rates are drawn from the uniform distribution $U[0, 0.6]$. Panel A shows the Laffer curve for unit taxes under a profit-weight assumption with $\kappa = 0.5$. Panel B shows the Laffer curve for ad valorem taxes under a Bertrand assumption. The neural network used in estimation is medium-sized, includes demand derivatives, and is trained on $T = 1,000$ markets.

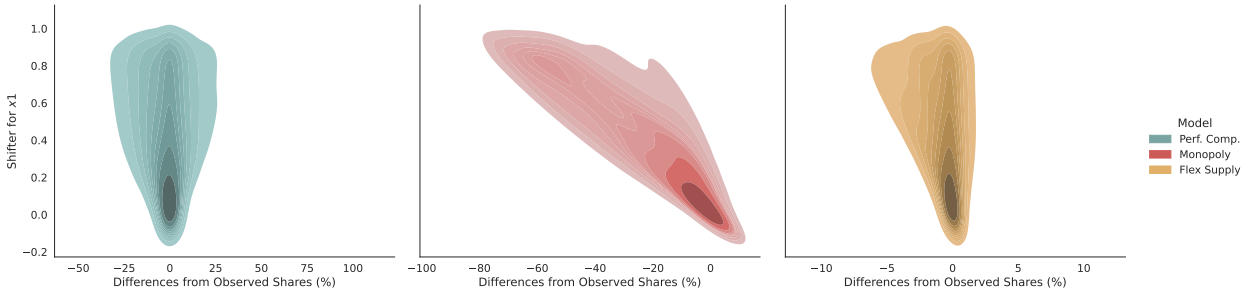
FIGURE G6: Laffer Curves for Unit and Ad Valorem Taxes



	A. Bertrand (Unit)	B. Profit-Weight (Unit)	C. Bertrand (AV)	D. Profit-Weight (AV)
Bertrand	-	0.51	-	1.02
Monopoly	2.07	0.99	3.28	3.82
Perf. Comp.	0.75	1.19	2.12	2.41
Flex Supply	0.13	0.06	0.17	0.51

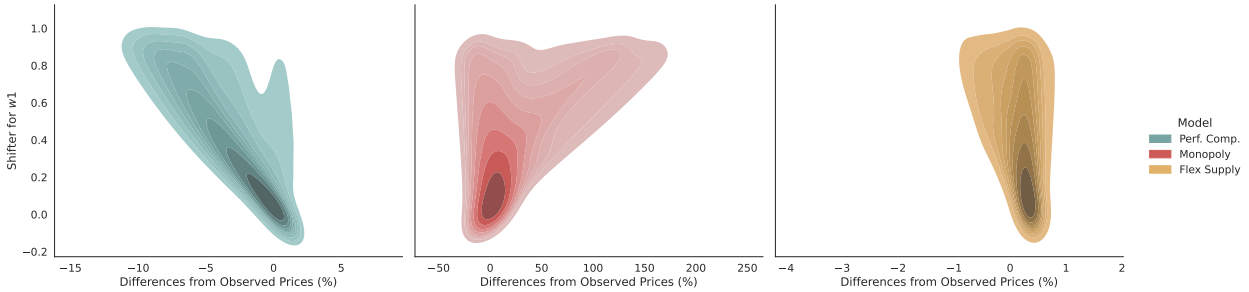
Notes: The figure displays Laffer curves for the flexible model estimated with VMM and a set of standard models under different conduct assumptions. This set of simulations limits the variation available in tax rates. Specifically, unit taxes are drawn from the uniform distribution $U[1, 2]$ and ad valorem tax rates are drawn from the uniform distribution $U[0, 0.2]$. Panel A shows the Laffer curve for unit taxes under a Bertrand assumption. Panel B shows the Laffer curve for unit taxes under a profit-weight assumption with $\kappa = 0.5$. Panel C shows the Laffer curve for ad valorem taxes under a Bertrand assumption. Panel D shows the Laffer curve for ad valorem taxes under a profit-weight assumption with $\kappa = 0.5$. The neural network used in estimation is medium-sized, includes demand derivatives, and is trained on $T = 1,000$ markets.

FIGURE G7: Product Characteristics Regulation, Bertrand DGP



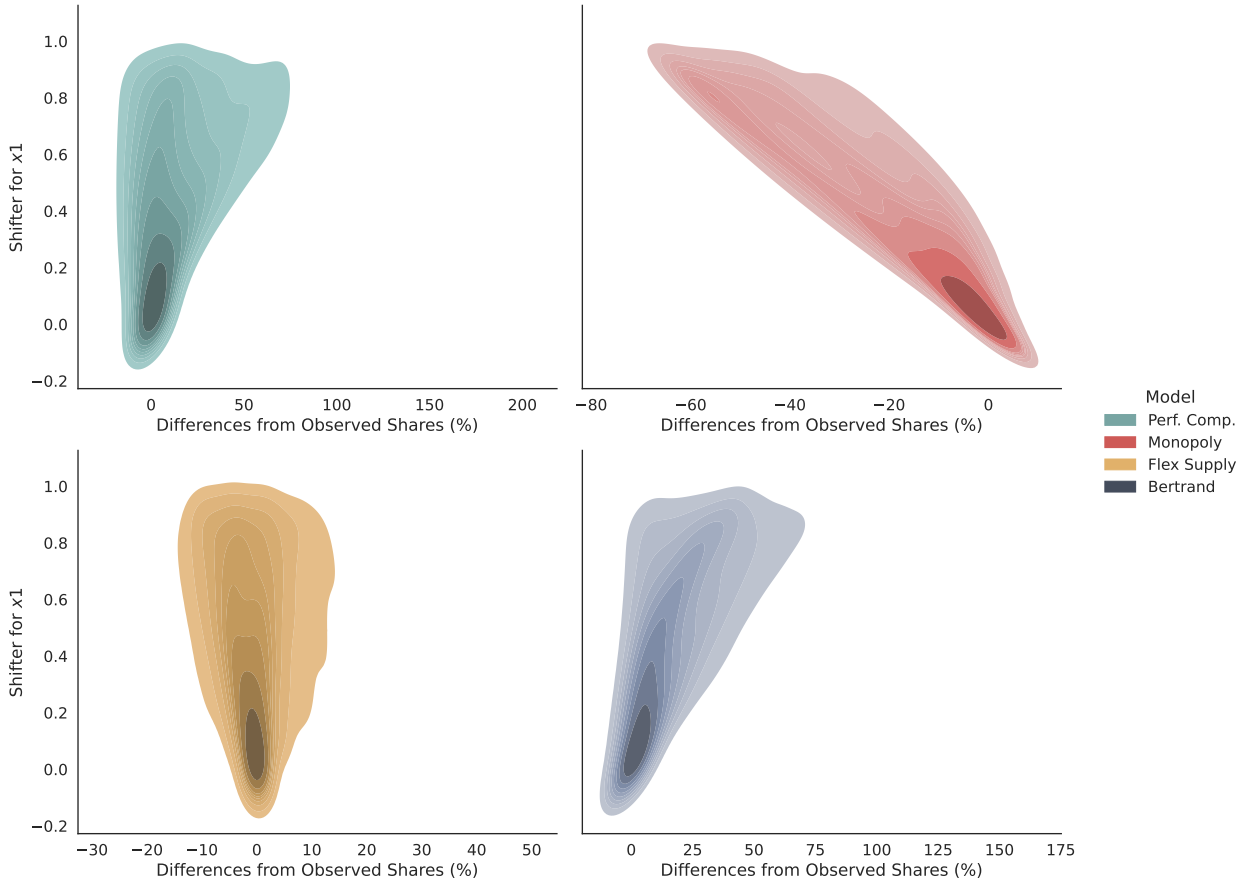
Notes: The figure displays counterfactuals in which product characteristics change for the Bertrand DGP. The regulation increases x_1 by varying factors. The neural network used in estimation is small with a 3×3 hidden layer and $T = 10,000$ markets.

FIGURE G8: Cost Shifter Regulation, Bertrand DGP



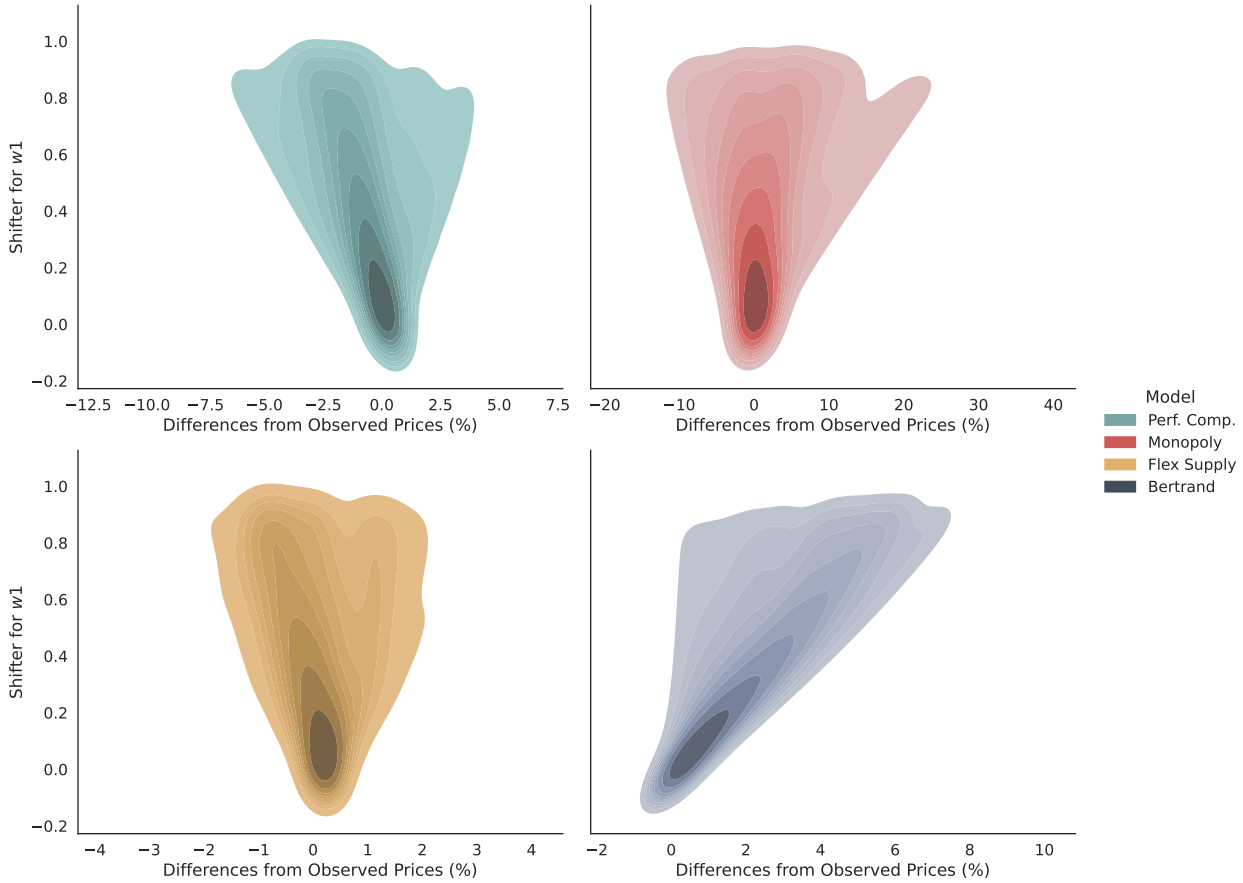
Notes: The figure displays counterfactuals in which product characteristics change for the Bertrand DGP. The regulation increases w_1 by varying factors. The neural network used in estimation is small with a 3×3 hidden layer and $T = 10,000$ markets.

FIGURE G9: Product Characteristics Regulation, Profit-Weight DGP



Notes: The figure displays counterfactuals in which product characteristics change for the profit-weight DGP ($\kappa = 0.5$). The regulation increases x_1 by varying factors. The neural network used in estimation is small with a 3×3 hidden layer and $T = 10,000$ markets.

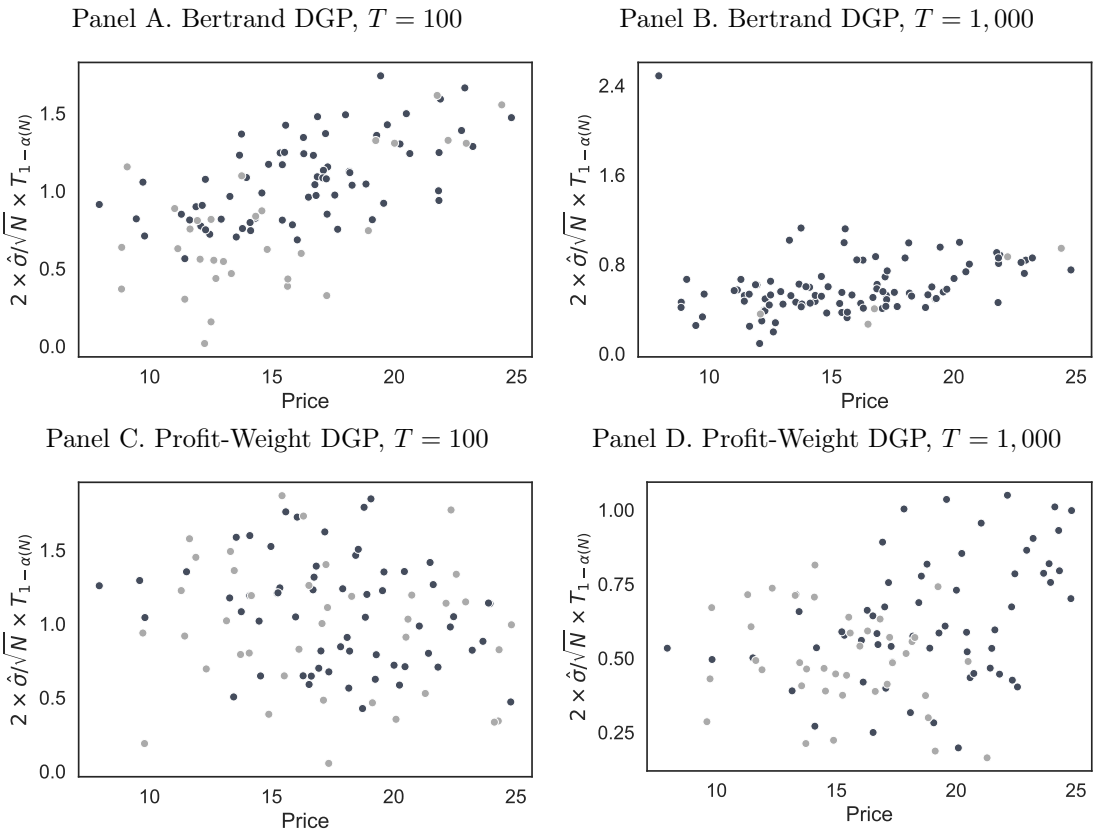
FIGURE G10: Cost Shifter Regulation, Profit-Weight DGP



Notes: The figure displays counterfactuals in which cost shifters change for the profit-weight DGP ($\kappa = 0.5$). The regulation increases x_1 by varying factors. The neural network used in estimation is small with a 3×3 hidden layer and $T = 10,000$ markets.

G.3 Additional Inference Results

FIGURE G11: Inference on Counterfactual Merger Simulation Prices



	$T = 100$		$T = 1,000$	
	A. Bertrand	B. Profit-Weight	C. Bertrand	D. Profit-Weight
Within	69%	61%	95%	61%
Outside	31%	39%	5%	39%

Figure displays the width of confidence intervals for prices resulting from a counterfactual product exit under a Bonferroni correction. Panels A and B show the results under Bertrand conduct and constant costs with $T = 100$ and $T = 1,000$, respectively. Panels C and D show the results under profit-weight conduct and economies of scale with $T = 100$ and $T = 1,000$, respectively. The neural networks used in estimation are medium with 20×20 hidden layers and exclude demand derivatives.

H Additional Empirical Results

H.1 Data Construction

We construct a database of the US airline industry from 2005-2019. We obtained a quarterly random 10 percent sample of purchased airline tickets from the well-known Airline Origin and Destination Survey (DB1B) database released by the US Department of Transportation. Following Azar et al. (2018) and Kennedy et al. (2017), a market is defined as a pair of cities, regardless of the flight direction. We match cities to Metropolitan Statistical Areas and collect data on the populations of these MSAs from the Bureau of Economic Analysis. A product is a one-way trip that services a particular city-pair and is defined at the carrier-market-quarter level. Market sizes are measured as the geometric mean of the origin-destination endpoint populations

H.1.1 Sample Selection

We exclude markets with fewer than 20 passengers per day, as airline behavior on these thin, possibly seasonal, routes is unlikely to represent normal competitive behavior in the industry. We also drop itineraries with a ticket carrier change at the connecting airport since these tickets cannot be assigned to a unique ticketing carrier. Finally, we drop every ticket with a fare lower than $25 and higher than $2,500 since these tickets are likely the result of reporting errors.

For each carrier-market-quarter, we begin by calculating the product's average price, total passengers, and average distance. Additionally, we construct each product's extra miles - the difference between the average distance in miles and the nonstop distance in the market - and the fraction of nonstop tickets sold. Averages are weighted by the number of passengers. We remove any products with fewer than 800 quarterly passengers. These products arguably have a weak impact on the competitive behavior of carriers with higher market shares, and although this is standard practice in this literature (e.g., Berry and Jia, 2010) because the dimensionality of the input space is very important in our application, this rule allows us to find more markets with effectively fewer carriers. Summary statistics for the analysis sample are presented in Table H1.

TABLE H1: Summary Statistics

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Average Fare	216.782	82.819	25.000	169.774	209.072	252.704	2,492.017
Total Passengers	352.395	1,184.007	1	19	69	210	70,909
Average Distance	1,386.534	688.694	67.000	861.000	1,255.081	1,872.045	7,731.500
Average Nonstop Miles	1,170.974	618.988	67.000	678.000	1,035.000	1,600.000	2,783.000
Average Extra Miles	215.560	247.441	−1.000	37.548	136.000	305.862	5,118.000
Share Nonstop	0.205	0.363	0.000	0.000	0.000	0.179	1.000
Origin Hub	0.153	0.360	0	0	0	0	1
Dest. Vacation	0.096	0.294	0	0	0	0	1
LCC	0.207	0.405	0	0	0	0	1
Major	0.920	0.271	0	1	1	1	1
Legacy	0.733	0.443	0	0	1	1	1
Presence	0.148	0.142	0.00000	0.050	0.099	0.201	1.000
Num Markets	50.434	30.156	1	27	47	72	146
Share	0.001	0.003	0.00000	0.0001	0.0004	0.001	0.095
Within Share	0.184	0.219	0.00001	0.022	0.096	0.268	1.000

Notes: The table reports summary statistics for our analysis sample. We include the mean, standard deviation, 25th percentile, median, 75th percentile, and max of each of the variables included in our sample.

H.2 Demand Estimation

We include additional details on demand estimation introduced in Section 6.2. In our demand system, consumer i receives utility from product j in market t with the following indirect utility:

$$u_{ijt} = \alpha p_{jt} + x_{jt}\beta + \xi_{m(t)} + \xi_{jt} + \zeta_{it} + (1-\rho)\varepsilon_{ijt}$$

The vector x_{jt} includes the share of nonstop flights, average distance in thousands of miles, the squared term of average distance in thousands of miles, and the logged number of fringe firms (plus one to avoid zero issues). The last term is included to focus on the demand for major carriers while controlling for additional variation over time in market structure across origin-destination pairs. The term $\xi_{m(t)}$ is a set of origin-destination fixed effects. ξ_{jt} and $\zeta_{it} + \varepsilon_{ijt}$ are unobservable shocks at the product-market and individual-product-market levels, respectively. We assume that ε_{ijt} is distributed Type I Extreme Value and ζ_{it} is distributed according to the conjugate distribution (Cardell, 1997). We close the model by normalizing the utility of consumer i from the outside option to $u_{i0t} = \varepsilon_{i0t}$. Given the structure of utility and distributional assumptions, market shares

s_{jt} are a function of observables, unobservables, and parameters in the standard form from Berry et al. (1995).

The identifying assumption for demand is that the moment condition $\mathbb{E}[\xi_{jt} z_{jt}^D] = 0$ holds for a vector of demand instruments z_{jt}^D. Following Berry et al. (1995), we include the average rival distance, the average number of markets a rival serves, and the number of rival carriers. The last instrument is especially useful for identifying the nesting parameter. The results of demand estimation are presented in Table H2. The results and median own-price elasticities are in line with the literature.

TABLE H2: Demand Estimates

	$\log(s_{jt})$ - $\log(s_{0t})$
Average Fare	-0.0048***
	(0.0004)
$\log(S_t)$	0.8356***
	(0.0133)
Share Nonstop	0.4030***
	(0.0282)
Average Distance (1,000's)	-0.4881***
	(0.0498)
Average Distance^2 (1,000's)	0.0485***
	(0.0045)
log(1 + Num. Fringe)	-0.2642***
	(0.0057)
R^2	0.94238
Observations	1,283,472
Own-price elasticity	-5.1652
Origin-destination fixed effects	✓

Notes: The table presents results from demand estimation. Prices and log-within-shares $\log(S_t)$ are instrumented with the average rival distance, the average number of markets a rival serves, and the number of rival carriers. We include origin-destination fixed effects and cluster standard errors at the origin-destination level. The median own-price elasticity is -5.1652 which is in line with the literature.

H.3 Supply Estimation

In Section 6.2, we briefly introduced the supply side and focused on the flexible markup function. In the Bertrand specification, we assume that inferred marginal costs are linear in observable cost shifters:

$$p_{jt} - \Delta^B_{jt} \equiv c_{jt} = w_{jt}\gamma + \Gamma_{m(t)} + \omega_{jt}$$

We include only the average distance in thousands of miles in w_{jt} and origin-destination fixed effects as $\Gamma_{m(t)}$. The results are presented in Table H3. We find that distance is positively associated with marginal costs inferred under the Bertrand assumption of conduct.

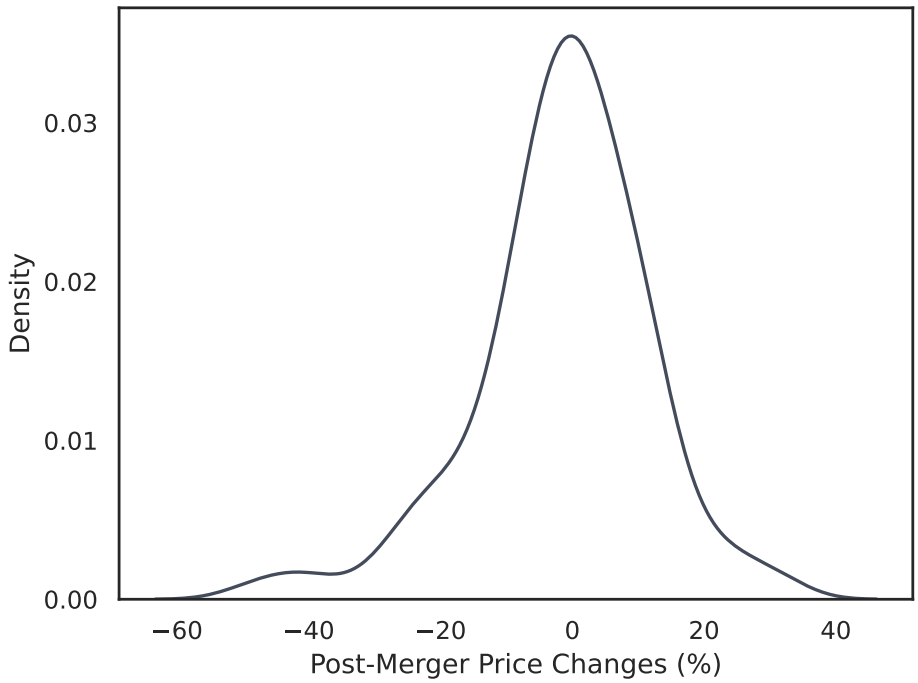
TABLE H3: Bertrand-Implied Marginal Cost Estimates

	Marginal Cost
Average Distance (1,000's)	63.17***
	(0.9502)
R^2	0.42757
Observations	1,283,472
Origin-destination fixed effects	✓

Notes: The table presents results from supply estimation. We include origin-destination fixed effects and cluster standard errors at the origin-destination level.

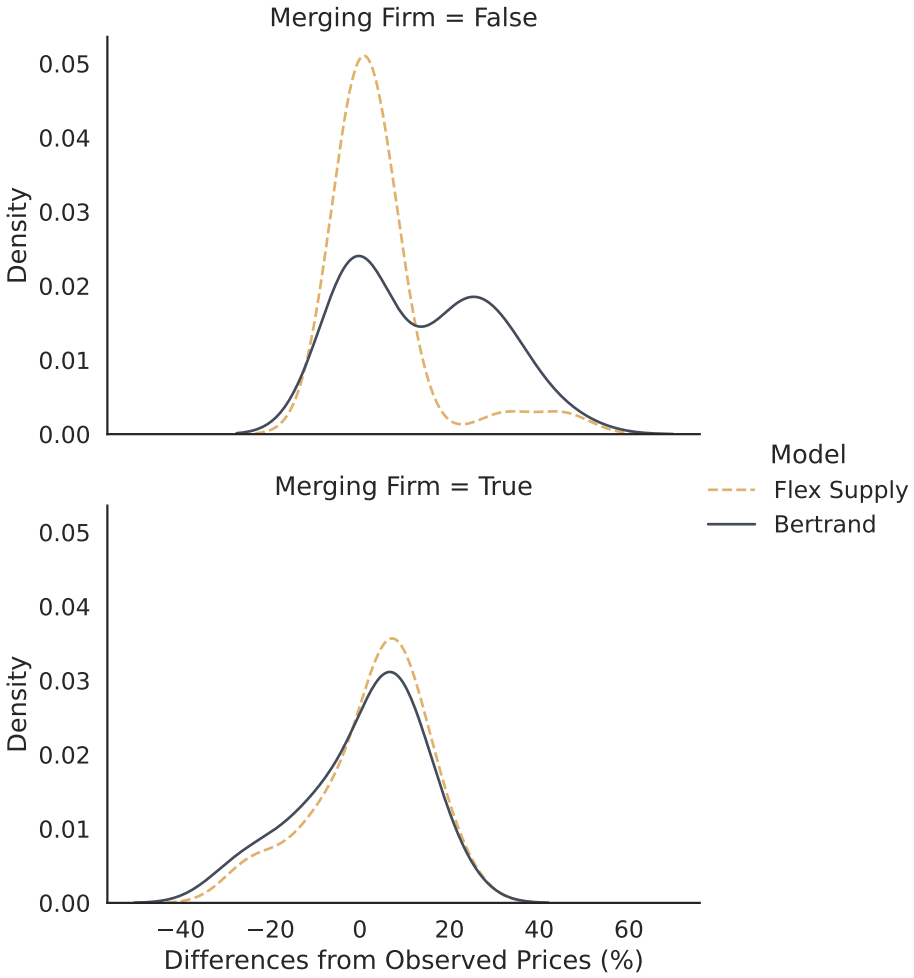
H.4 Counterfactuals

FIGURE H1: Price Change Distribution



Notes: The figure plots the observed post-merger price changes after the US-AA merger.

FIGURE H2: Post-Merger Price Prediction Error by Status



Notes: The figure reports merger simulation results for the flexible model estimated with VMM (in yellow) and the standard merger simulation model (in blue) broken down by firm status. It reports the distribution of percent differences between observed and post-merger predicted prices.

I Using VMM

Researchers can easily use our implementation of VMM in simulations and empirical applications. The models are implemented using the `torch` package in Python. Compute times are manageable by modern standards. Our simulations for $T = 100$ finish within a few minutes; $T = 1,000$ within an hour; and $T = 10,000$ in about a day. The model for our empirical application runs in about 16 hours. We build an easy-to-use wrapper around the machinery of VMM. Researchers specify cost shifters, product characteristics, and instruments they would like to include. We allow the user to optionally include demand derivatives, which we build in the background.

```
# Import packages
import pandas as pd
import pyblp
import vmm

# Load data, set variables
data = pd.read_csv('data.csv')
exog, char, ins = ['w1'], ['x1', 'x2'], ['z1', 'z2', 'z3']

# Estimate demand
demand = pyblp.Problem(pyblp.Formulation('1 + prices + x1 + x2'), data).solve()

# Set up scenario
scenario = Supply(exog=exog, char=char, ins=ins, derivatives='yes')
scenario.setup(data, demand)

# Fit model
problem = SupplyProblem(scenario, h_dim=3)
results = problem.solve()

# Counterfactuals
divest(demand, scenario, results, firms=[0])
merger(demand, scenario, results, firms=[[0, 1]])
laffer(demand, scenario, results)
regulation(demand, scenario, results, ['x1', 0.5])
regulation(demand, scenario, results, ['w1', 0.5])
```