Structural Deep Learning*

Max H. Farrell† Tengyuan Liang‡ Sanjog Misra§

Preliminary Draft - Please do not Circulate
November 16, 2024

Abstract

This paper introduces a novel approach that integrates deep neural networks (DNNs) into structural economic models to enhance flexibility and accuracy of estimation while preserving interpretability. By combining machine learning techniques with economic theory, we address the limitations of both approaches: DNNs provide the capacity to learn complex, non-linear relationships, while the structural model ensures the estimates remain interpretable and suitable for economic decision-making and policy analysis. Our approach extends a standard parametric model by enriching its parameters into fully flexible functions of observed covariates, estimated using a structured DNN architecture. We apply this methodology to demand estimation in consumer choice contexts, demonstrating how personalized pricing strategies can be derived and evaluated using double machine learning for valid inference. Additionally, we provide theoretical results on the non-asymptotic bounds and convergence rates of our structural deep learning estimators, as well as conditions for valid inference using influence functions.

Keywords: Deep Learning, Structural Modeling, Heterogeneity, Machine Learning, Influence Function, Neyman Orthogonality, Semiparametric Inference

*This paper replaces "Deep Learning for Individual Heterogeneity: An Automatic Inference Framework". We thank Whitney K. Newey for helpful comments and suggestions. Janani Sekar and Kirill Skobelev provided outstanding research assistance. We also thank Thomas Weimann for assistance with the brand choice application section. Replication files are available upon request.

†Department of Economics, UC Santa Barbara
‡Booth School of Business, University of Chicago
§Booth School of Business, University of Chicago

1 Introduction

There has been a recent influx of machine learning in economic research. This has been driven by the widespread availability of big data and the recent development of tools that can exploit such data to great effect. With the dramatic advances in computing power and optimization methods, machine learning has become available to all researchers. The expressive power of machine learning has lead to the temptation to "let the data speak" and avoid imposing any restrictions or assumptions on numerical analysis, assuming that all relevant features will be picked up from the data. We argue that actually economic structure and machine learning are complements in modeling, not substitutes. Both have their strengths and weaknesses, but moreover, by combining them appropriately one can use the strengths of each to remedy the shortcomings of the other.

Economic models are designed to obey theories and facts from the discipline. This means that, when taken to data, the estimates and inferences from structural models are (economically) interpretable and useful for decision and policy making. Crucially in many economic decision problems, the structure of the model, and the discipline it imposes, allows for interpolation and extrapolation, and thus optimization. Models without economic structure can be useful descriptively, but because the only structure comes from statistical assumptions, they cannot be reliably used in economic optimization for data-driven choices of pricing, production, or other strategies.

However, economic theories, and hence the structural models based upon them, are incomplete in important ways for data analysis. Here statistical and machine learning methods can be exploited. Theory may not dictate things like functional form or the presence of interaction effects. Such elements of the model are unspecified by theory but crucial for data analysis, and must be learned accurately as they play a role in the ultimate decision. One leading example is heterogeneity. Even in cases where the presence of heterogeneity is implied economic reasoning, which is not a given, the form and type will not come from theory and must be learned from data. Seasonality, or other time patterns, are another commonly encountered phenomenon which is known to exist and unspecified by theory. Production (or consumption) will change over time and capturing this pattern in the data is crucial for analysis and decision making. Correlation structures, covariate importance, and effect spillover patterns are other examples where a mix of theory and flexilibity is called for.

Flexibly learning patterns and heterogeneity is precisely the strength of modern machine learning

(ML) methods. ML often cannot learn economic structure, because such methods are explicitly designed for flexibility, while structure by its nature means constraints. With complete flexibility, the results of ML are often not interpretable economically and cannot be used for optimization, which requires inter- and extra-polation. Pure ML extrapolates based on statistical structure built into the method, not on ecomics. This phenomenon is shown in a simple empirical example in the next section. Thus while ML cannot replace structure, it can useful augment structure.

We show how to embed machine learning, in the form of modern neural networks, into economic modeling. Our approach begins with a structural model imposed by the researcher. The model relates the outcomes $\boldsymbol{y}_i$ to the covariates of interest $\boldsymbol{t}_i$ (treatments or policy-relevant variables) and depends on parameters of interest $\boldsymbol{\theta}$, which are to be estimated from the data. We suppose this model is captured by a loss function $\ell(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{\theta})$. The structural model encodes restrictions and constraints on the data generating process and the parameters $\boldsymbol{\theta}$ have direct economic interpretations. From an econometric point of view, is a standard parametric estimation and inference problem.

We enrich the model by changing the parameters $\boldsymbol{\theta}$ to *parameter functions* $\boldsymbol{\theta}(\boldsymbol{x})$, which are fully flexible functions of a vector of observed characteristics $\boldsymbol{x}_i$. The new model is then $\ell(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{\theta}(\boldsymbol{x}_i))$. This changes adds flexibility but maintains all the structure of the original model. The incompleteness of theory is reflected in the flexibility of treating $\boldsymbol{\theta}(\boldsymbol{x}_i)$ as a fully unknown function. This is exactly where the strength of machine learning is exploited, and how the ML complements the economics. Either one alone is insufficient: structure without machine learning would lose the richness and fail to capture important patterns in the data, while naively applying machine learning without structure would lose the interpretability and the utility in policy/decision making.

We estimate the parameter functions $\boldsymbol{\theta}(\boldsymbol{x}_i)$ using deep neural networks. Neural networks allow the structure of the model to be directly encoded into the estimation through the network architecture. The required change to the architecture is intuitive and simple to implement, requiring only a few extra lines of code. The power and flexibility of deep learning is directed through the architecture to learn the parameter functions $\boldsymbol{\theta}(\boldsymbol{x}_i)$. This is in contrast to standard deep learning implementations (and much of ML in general) that are focused entirely on prediction or classification tasks (learning regression functions) whereas our method allows for learning any economically interesting parameter, such as coefficients, variances, elasticities, et cetera, all as rich, heterogeneous functions.

By exploiting the fact that we enrich a parametric model, we also obtain valid two-step semipara-

metric inference for finite-dimensional parameters using an insight into the influence function in such context. We show that ordinary derivatives can continue to characterize the influence function in the semiparametric case, just as in the parametric case. Further, the only thing required is the value of these derivatives at the data points, not as functions in general. This means we can deliver an influence function in a very wide variety of settings, and so we can obtain valid inference using double machine learning in more contexts than previously possible. Availability of influence functions, or more generally, orthogonal scores, is a limiting factor in conducting inference after first-step ML. Further, and importantly for practice, the automatic differentiation engine used in neural network estimation can be exploited to automatically obtain the influence function, without any human derivations needed, only computation and nonparametric estimation.

We illustrate our method studying short-term loans using the data of Bertrand et al. (2010). We study the value of advertising characteristics, as in the original work, but in our case the model is fully flexible in heterogeneity. We also provide a demand application in the context of brand choice models using scanner panel data. We then model demand as a function of prices and marketing efforts, and once the demand function is estimated, we compute optimal prices as a function of characteristics (third degree price discrimination) and conduct inference on quantities such as counterfactual profits. This involves applying double machine learning ideas to objects not available in closed form (the solution to a fixed point problem), which illustrates the generality of our inference method.

Our main focus is on the methodology of using deep neural networks in economic modeling. However, we do justify each step theoretically. For the structured deep neural networks we prove nonasymptotic bounds (and implied convergence rates), building on Farrell, Liang, and Misra (2021). These rates depend only on the dimension of the heterogeneity $\boldsymbol{x}_i$, because the relationship of $\boldsymbol{t}_i$ to $\boldsymbol{y}_i$ is not learned from the data. Our bounds apply to many settings of interest, requiring only mild conditions on the loss function $\ell(\cdot)$ and standard smoothness requirements. For inference, the influence function derivation is the key contribution and is shown to be valid. We can then obtain asymptotic normality by applying the double machine learning methodology of Chernozhukov et al. (2018a).

The rest of the paper is organized as follows. We next review some related literature. Section 2 then shows a simple, motivating example demonstrating the value of structure. Our structured deep learning method is described in Section 3. Inference is discussed in Section 4. We conduct

two empirical applications of our our methods in Section 5.1. Section 6 summarizes the theoretical results. Finally, Section 7 concludes. The Appendix contains implementation details, discussion of related contexts, and all proofs. We also therein discuss the important and intuitive special case of generalized linear models in order to illustrate the ideas and connect to other results.

1.1 Related Literature

Our work touches on several areas of economics, econometrics, and statistics, to which we cannot hope to do justice in a few paragraphs. We will give an overview here, with some further discussion is given throughout the paper.

First, our work is related to the recent interest in integrating ML into economic research. Most, if not all, of this work has used ML to learn regression functions or make predictions, even if these feature in other economic contexts. ML is often a substitute for nonparametric regression, and succeeds where classical methods fail due to the complexity of the data. Often causal inference is the end goal, but our work, being motivated by structure, is more related to other economic contexts. See Varian (2014) and Bajari et al. (2015) for reviews.

Our structured deep learning approach, and the accompanying nonasymptotic bounds, connect our work to the recent literature studying the statistical properties of ML methods. Our theoretical results build directly on Farrell, Liang, and Misra (2021). Conceptually, Athey, Tibshirani, and Wager (2019) and Foster and Syrgkanis (2023) may be the most closely related, as both focus on quantities other than prediction functions, use models beyond regression, and utilize orthogonal scores as key ingredients. The contexts and results are different, as we focus on deep learning and semiparametric inference. Athey, Tibshirani, and Wager (2019) study random forests and nonparametric inference at a point, while Foster and Syrgkanis (2023) are concerned with risk bounds under weak conditions.

The use of influence functions explicitly with the goal of obtaining valid inference under weaker conditions is not new, whether for classical nonparametrics (Cattaneo, 2010) or machine learning (Farrell, 2015). Our work contributes most directly to the recent literature applying the double machine learning (DML) method of Chernozhukov et al. (2018a). DML requires an orthogonal score, which an influence function provides, and thus many papers stick to known examples or hand-derive an influence function. Our contribution here is the derivation of a generic influence

function that covers a broad class of models and can be computed automatically. We share this goal with the "auto-DML" method (see Remark 3); our approach is different and has relative strengths and weaknesses.

2 The Value of Structure

The value of structure in economic modeling has been well understood for decades. Indeed, the half-century old Lucas Critique is a (major) example of the value of structural thinking. With the power of ML, there is a temptation to learn everything from data alone. This view has its origins in the machine learning literature, which has focused exclusively on prediction, but has seeped into other contexts. Zadrozny (2004) writes, in the context of classification, that the ML community is "interested in the predictive performance of the model and not in making conclusions about the underlying mechanisms that generate the data." So successful is machine learning, and deep learning in particular, at these prediction tasks, that it can be tempting to assume that any problem can be tackled with accurate enough predictions. In some cases, such as hand-writing recognition, pure prediction is enough. But economic decision making cannot proceed without these "underlying mechanisms", that is, without structure and economics. The danger in naively applying ML in economic decision making contexts is that economic assumptions and structure are replaced with statistical assumptions and the structure of estimators.

To demonstrate this danger, we use the data of Bertrand et al. (2010), which is analyzed more fully in Section 5.1. Here our goal is to estimate the demand for loans as a function of the interest rate, and then to find the optimal interest rate offer. This is straightforward demand estimation, where the price is the interest rate and the good being "purchased" is the loan.

For simplicity we use only the second two waves of Bertrand et al. (2010) and only high risk individuals, leaving us with 40,000 observations, a binary outcome, and with 46 values of the interest rate, a continuous predictor. Moreover, this is an experiment, so the price variable has been randomized, so endogeneity is not a concern. Thus, our data is ideal for this use, both economically and statistically. This is a tailor-made classification task, and the naive ML view would hold that if conversion for any interest rate offered can be predicted accurately, then the interest rate can be optimized. However, as we will see, neglecting the "underlying mechanisms that generate the data"

will lead to a failure.

Figure 1: **Estimated Demand Functions.** Bertrand et al. (2010) asdfasd

(a) *Random Forests* (b) *Neural Network* (c) *Structural Logit*

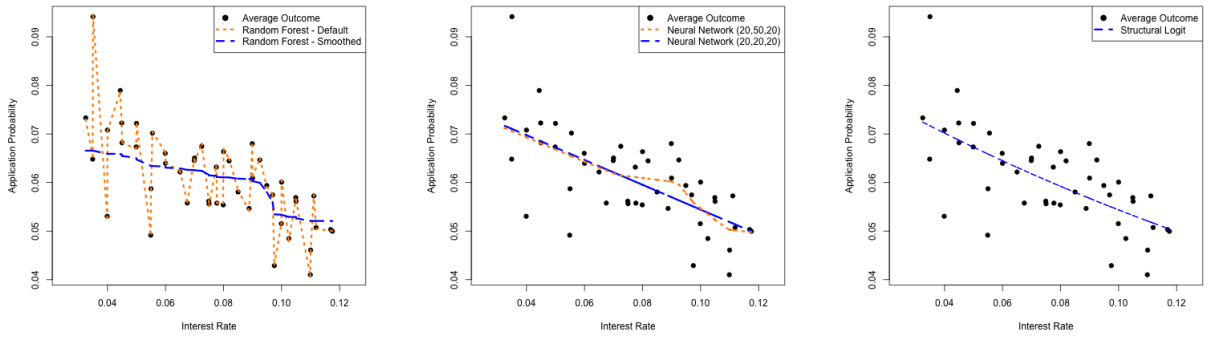


Figure 1 shows three different approaches to estimating the demand function in this data. In each case, the dots show the empirical application rate (average purchasing decision) at each offered interest rate. Panel (a) shows two random forest estimates using the `ranger` package. The dotted orange line uses the default settings. This appears to be undersmoothed and thus the blue dashed line forces a smoother fit using an ad-hoc restriction on the tree depth. It is easy to (by eye) reject the default random forest as an unreasonable estimate of the demand function, although it should be noted that even this judgement reflects economic structure, not statistical/ML criteria: nothing in machine learning says the curve should be smooth or downward-sloping. Panel (b) shows two neural network fits, the blue dashed line is three layers with 20 nodes per layer and the dotted orange line adds richness by using 50 nodes in the middle layer. Finally, panel (c) shows the fit from a linear randomly utility model with a logistic error distribution, the basic workhorse structural model in IO.

Both the smoothed-out forest and both neural networks produce curves that could reasonably be demand functions. However, as we continue with the pricing problem, it will become clear how different these fits really are. Figure 2 takes the smoothed forest, (20,20) neural network, and binary choice logit and extrapolates these estimated demand functions from the support of the data (3.25% to 11.75%) to, in panel (a), zero to 20%, and in panel (b), for all interest rates up to 200%.

The forest is the most striking, with completely flat demand for any higher interest rate. This is mechanical and generic: a forest is an average of trees, which are piecewise flat, and so at the edge of the support the forest fit must be flat, hence a flat estimate at all higher values, in any data set. On the support of the data, the neural network fit appeared quite linear and very similar to the

logistic model. However, when we extrapolate, we see how different the fit is, due to the invisible (when looking at the sample) complexities and nonlinearities of the neural network estimate.

Figure 2: **Extrapolated Demand Functions.** Bertrand et al. (2010) caption caption caption

(a) *Extrapolating from zero to 20%* (b) *Extrapolating from zero to 200%*

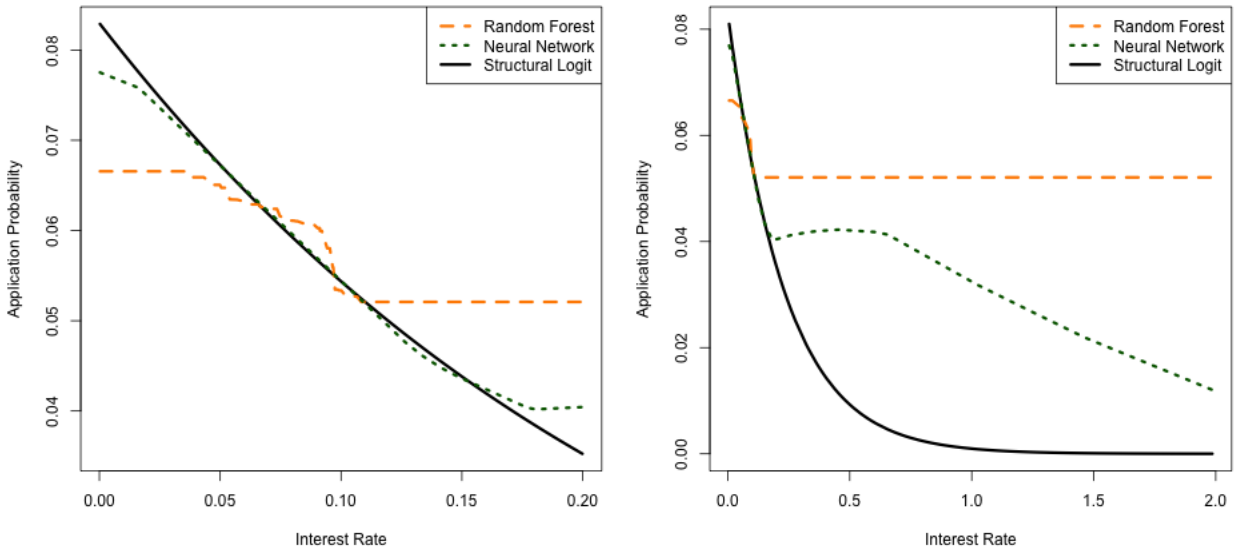


Figure 3 shows the revenue expected from each demand function. While this does not include costs, it is clear what each estimate implies for an optimal price. The structural model gives well-behaved demand and revenue curves and yields a reasonable optimal price. With flat demand, the forest revenue grows without bound, yielding infinite prices. Again, this is mechanical and generic. The neural network at least gives a revenue curve that eventually declines, but only for very high rates. The difference here again stands in stark contrast to the similarity of Figures 1(b) and (c). One can dismiss these findings with the argument that no reasonable decision maker would ever think infinite prices is appropriate, but again, this using economic structure to inform the ML, not the other way around.

These phenomena are not related to statistical uncertainty. Figure 5 shows a 95% confidence band for the demand function using Cattaneo et al. (2023). The band contains all three estimates and any manner of other possibilities. Different fits would yield different optima (except the forest, which always yields infinite prices, and thus has zero uncertainty) but the issues will persist, and potentially be exacerbated, for any realization of the data.

This simple example shows that, even in a data set that is in many ways ideal for the ML

Figure 3: **Implied Revenue Functions.** Bertrand et al. (2010) caption caption caption

Figure 4: *Implied Revenue*

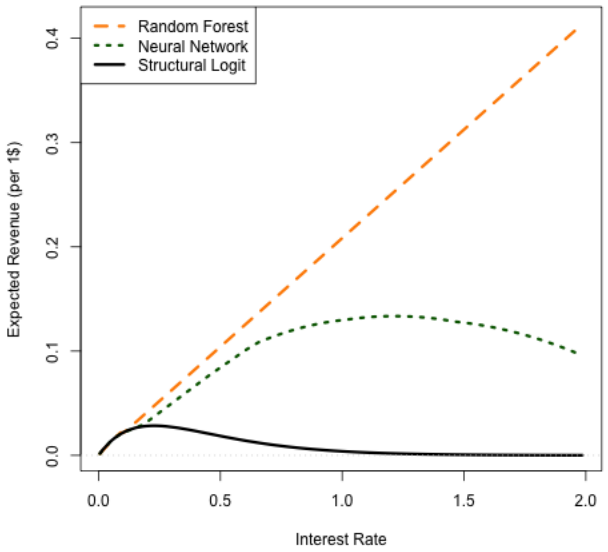
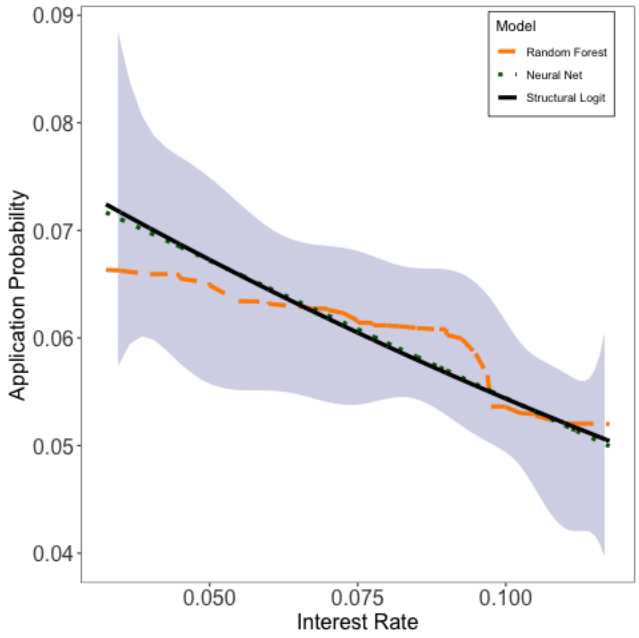


Figure 5: **confidence band using Cattaneo et al. (2023)** Bertrand et al. (2010) caption caption caption

Figure 6: *Confidence Band*



approach, economic structure cannot ignored. Both the forest and the neural network impose their own structure on the data, but this structure is opaque and is based on data-fitting concerns, not economic decision making. Decision making necessarily depends on inter- and extrapolation, which

by definition is based on assumptions and not upon the data. These assumptions can be economic or statistical. This example shows that this statistical structure may not be appropriate for economic problems.

Finally, it is worth noting that structure and structural reasoning, though it goes by other names, is commonly used in machine learning problems. Convolutional neural networks are a particular structure that is superior for image recognition. The transformer architecture (Vaswani et al., 2017) is behind the recent success of large language models. Structure at its core means constraints and restrictions, and these popular architectures are no different.

3 Embedding Deep Learning in Structural Models

We show that ML, in the form of deep learning, can be embedded into structural modeling in order to capture the benefits of both. In this section we describe our structured deep learning approach and describe the two step estimation and inference problem. Inference is treated in the following section. Theoretical results are postponed to Section 6.

The starting point is a standard parametric structural model that encodes the researcher's restrictions on how the outcomes $\boldsymbol{Y} \in \mathbb{R}^{d_{\boldsymbol{Y}}}$ relate to the variables of interest $\boldsymbol{T} \in \mathbb{R}^{d_{\boldsymbol{T}}}$.[1] These "treatment" variables can be randomized or not and can be continuous, discrete, or mixed. What is important is that these are the variables of interest and it is on the relationship of $\boldsymbol{Y}$ to $\boldsymbol{T}$ that the researcher places structure. The restrictions of the model are encoded in a loss function $\ell(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{\theta})$, which depends parameters $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$ to be estimated from the data. The true parameters minimize the expected loss:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min}\, \mathbb{E}\left[\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta})\right], \tag{3.1}$$

where $\boldsymbol{\Theta} \subset \mathbb{R}^{d_{\boldsymbol{\theta}}}$ is the appropriate parameter space. Inference is then conducted a second stage object that is an average of a known function $\boldsymbol{H}$, that depends on the parameters, covariates, and possibly some fixed value of interest $\boldsymbol{t}^*$, as in

$$\mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{t}^*)], \tag{3.2}$$

[1] *Notation.* Vectors and matrices will be written in boldface. Capital letters are used for population random variables; lower case for realizations. The expectation operator with respect to the true data generating process is denoted $\mathbb{E}[\cdot]$. The L_2 norm for a function $g(\boldsymbol{x})$ is $\|g\|_2 = \mathbb{E}[g(\boldsymbol{X})^2]^{1/2}$.

which include the parameters themselves (for $\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{t}^*) = \boldsymbol{\theta}$) and quantities such as counterfactuals and average marginal effects. Making $\boldsymbol{t}^*$ explicit notationally is useful for targeting and optimizing strategies, but conceptually could be included in the definition of $\boldsymbol{H}$. Other than restricting to per-observation losses and smooth functions (so that derivatives exist), the combination of Equations (3.1) and (3.2) encompass a wide variety of parametric models, including many M- and Z- estimation problems, such regression models, quasi/pseudo-likelihoods, or generalized estimating equations (Newey and McFadden, 1994).

The positives of structural modeling are widely appreciated. The model incorporates theory-based restrictions and constraints. This means that the estimated parameters, or transformations thereof, are interpretable and directly useful useful in policy analysis, decision making, and the formation of counterfactuals. The two-step nature of the problem reflects exactly these types of analyses, where second step counterfactuals are obtained after estimating primitives of the model in the first stage. It also means that the data is used more efficiently, since structure implies restrictions, which will be crucial when embedding machine learning or nonparametrics.

The major drawback of Eqn. (3.1) is rigidity: it does not allow for any heterogeneity in the relationship between $\boldsymbol{Y}$ and $\boldsymbol{T}$. If there is any such heterogeneity, which is almost a given in economics and social sciences, then the results will be biased and subsequent decisions or analyses will be erroneous. Concern over effect heterogeneity has been the focus of much recent research, particularly in program evaluation (Chernozhukov et al., 2018b; Wager and Athey, 2018; De Chaisemartin and d'Haultfoeuille, 2023). We may aim for being robust to heterogeneity, as in classical fixed effects models, or to capture and exploit heterogeneity, in more modern targeting contexts.

Our idea is to recast the parameters $\boldsymbol{\theta}$ as *parameter functions* $\boldsymbol{\theta}(\boldsymbol{X})$ that are fully flexible in observed covariates $\boldsymbol{X} \in \mathbb{R}^{d_{\boldsymbol{X}}}$, then estimate these functions with structured deep learning. Instead of the true model being the inflexible (3.1), we assume that

$$\boldsymbol{\theta}_0(\cdot) = \underset{\boldsymbol{\theta} \in \mathcal{F}}{\arg\min}\, \mathbb{E}\big[\ell\big(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X})\big)\big], \tag{3.3}$$

for a function class $\mathcal{F}$, which obeys standard restrictions detailed below (Assumption 2). The second

step parameter of interest is correspondingly enriched, to be

$$\boldsymbol{\mu}_0 = \mathbb{E}\big[\boldsymbol{H}\big(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}), \boldsymbol{t}^*\big)\big], \tag{3.4}$$

where we assume that $\boldsymbol{H}$ maps to $\mathbb{R}^{d_{\boldsymbol{\mu}}}$. We focus such $\boldsymbol{\mu}_0$ to keep exposition simple, but our method and results apply to GMM as well, as discussed in Appendix B.7.

Thus we have an enriched, ML-driven two step procedure that matches the logic of first estimating primitives and then counterfactual or policy analysis. Equation 3.3 is a specific, though quite general, formulation of nonparametric M-estimation (Gallant and Nychka, 1987), which together with (3.4) defines a broad class of semiparametric two-step estimation settings, matching the generality of (3.1) and (3.2). Our framework includes as special cases many of the standard semiparametric problems, including treatment effects, generalized (partially) linear models, and others. See Appendix B.

The important concept behind Eqn. (3.3) is that all the structure is maintained, but is enriched with fully flexible heterogeneity. The principles and constraints are still present and the parameters, now as parameter functions, are still directly interpretable and useful. $\boldsymbol{\theta}_0(\boldsymbol{x})$ is simply the effect for individuals with "type" $\boldsymbol{X} = \boldsymbol{x}$. Further, the $\boldsymbol{\theta}_0(\boldsymbol{x})$ are not "nuisance" functions. The view, particularly common in the realm of inference after ML, is that first step functions are literally a nuisance, i.e. something annoying that must be dealt with, but is not interesting. We object to this view: in many applications the learned heterogeneity is actually the most interesting part, and because the $\boldsymbol{\theta}_0(\boldsymbol{x})$ are interpretable functions in our framework, using them is simple.

We introduce a novel yet simple structural deep learning approach to estimate the parameter functions $\boldsymbol{\beta}_0(\boldsymbol{x})$. The key to this approach is the architecture is depicted in Figure 7. The expressive power of deep neural networks comes from the hidden layers. In standard approaches this flexibility is used to learn regression functions or form predictions, which determine both the target and the loss functions. In our approach, the hidden layers are directed through a *parameter layer* so that the power of ML is used entirely for learning the parameter functions. These then enter the *model layer* according to the structural model in order to optimize the loss. The neural network optimizes the parameter functions to minimize the same structural loss, not prediction loss. This formalizes exactly how machine learning and structure are complementary. The ML enriches the economic model, filling in gaps left by theory. Simultaneously the economics aids the implementation of ML

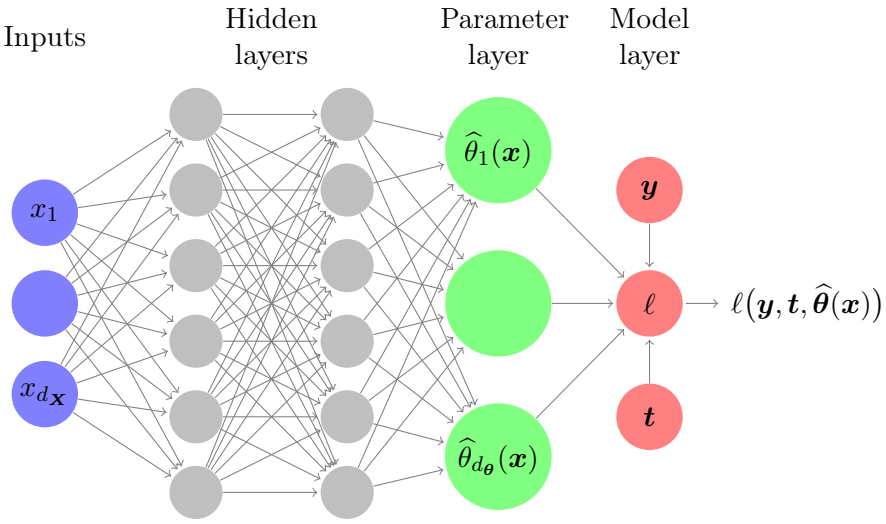


Figure 7: Structural deep learning. Figure shows a depicture of the deep neural network architecture for estimating the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ in the structural economic model $\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X}))$ defined in Eqn. (3.5).

to estimate structural objects with economic meaning and interpretation.

The estimates come from solving the empirical analogue of (3.3) where minimization is over a class of *structural* deep neural networks, i.e. those with architecture shown in Figure 7. We define

$$\widehat{\boldsymbol{\theta}}(\cdot) = \underset{\boldsymbol{\theta} \in \mathcal{F}_{\text{DNN}}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{\theta}(\boldsymbol{x}_i)), \tag{3.5}$$

where $\mathcal{F}_{\text{DNN}}$ is the class of deep neural networks that encodes not only the overall architecture but also tuning parameter choices of the width and depth of the network and the shape parameter given by the activation function. We use the standard fully connected feedforward neural network with ReLU activation for the hidden layers. To save space, we will not review deep learning basics here. Recent textbook treatments include: Goodfellow, Bengio, and Courville (2016) and James et al. (2021) for introductions and examples, Roberts, Yaida, and Hanin (2022) for theory, and Keydana (2023) for implementation. We discuss choice of the network parameters in our application in Section 5.1. Theorem 1 in Section 6 gives the convergence of these $\widehat{\boldsymbol{\theta}}(\cdot)$.

Our structural deep learning approach can be applied to any model/loss $\ell(\cdots)$ and is easy to implement. Only a few lines of code need be changed relative standard implementations, so that the flexibility of the network, represented by the hidden layers, is directed toward learning the parameter functions $\boldsymbol{\theta}_0(\boldsymbol{x})$ rather than the loss directly. We dub this "structural compatibility", meaning that

the structural model is easily baked into the estimation.[2] Neural network estimation relies on using gradients to find optima, proceeding by back propagation through the network. Our idea is simply to structure the final layers of this network, as shown in Figure 7, optimizing the loss through the parameter functions, instead of optimizing prediction directly, but the computation is the same. This intuition works the other way around also: in parametric structural estimation, $\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta})$ would be optimized in exactly the same way, but the gradients would stop at the parameter layer. Our idea simply extends this through additional layers to add nonparametric flexibility.

Of course, DNNs are not the only method that could be used to recover the parameter functions, nor do we claim any formal optimality property. Classical series estimators such as splines or polynomials are also structurally compatible, and this is one reason why they are a common method for nonparametric M estimation (see Gallant and Nychka (1987) for early work and Chen (2007) for a detailed theoretical treatment). Modern basis-function style methods, including the lasso and ridge regression, are also compatible. However, DNNs have several features which recommend them for structural estimation and inference in practice. First, deep learning is built on automatic differentiation, whic allows us to obtain the necessary influence function computationally for free for any $\boldsymbol{\mu}_0$ from (3.4) based on any first step (3.3). See Section 4. Second, DNNs handle discrete covariates seamlessly. While in theory including discrete covariates does not impact the convergence rate for many nonparametric estimators, obtaining these estimates in practice typically requires special care and custom methods (Racine and Li, 2004; Ma, Racine, and Yang, 2015). DNNs need no such customization, the inputs shown in Figures 7 can be any mix of continuous or discrete, and the rate still reflects only the number of continuous covariates. Finally, classical nonparametric methods often do not work for modern applications with more than several covariates.

The economic model holds globally, and we wish to match this in estimation, as it is important to learn counterfactual quantities in the second step. However, it is possible to use "local" methods to learn the value of the parameter functions at a point, i.e. $\boldsymbol{\theta}_0(\boldsymbol{x})$ for some $\boldsymbol{x}$, but often more complex and less transparent. Fan and Zhang (2008) discuss kernels and local polynomials while Zeileis, Hothorn, and Hornik (2008), Athey, Tibshirani, and Wager (2019), and Chatla and Shmueli

[2] In our coinage, "structural compatibility" is akin to the idea of a "global" smoother, typified by a series estimator. This use of "global" means that the estimator imposes a global smoothing structure, which is contrasted with "local" estimators that impose restrictions only locally, as in typical kernel regression. However, there is an ambiguity because some series estimators, such as splines or partitioning, are also "local" in the sense that only data near to the point of interest is used in the esstimation (Cattaneo and Farrell, 2013; Cattaneo, Farrell, and Feng, 2020).

(2020) use trees and random forests; all share our goal of learning non-prediction function. Also Chen (2007) views neural networks as sieves for generic nonparametic M estimation, as do we. Athey, Tibshirani, and Wager (2019) and Foster and Syrgkanis (2023) are perhaps the most recent closest antecedents to our work, as both move ML away from prediction and both use orthogonal scores as a key ingredient. Athey, Tibshirani, and Wager (2019) study random forests in a similar class of problems to (3.3). Their goal is inference on the nonparametric object $\boldsymbol{\theta}_0(\boldsymbol{x})$ at a point $\boldsymbol{X} = \boldsymbol{x}$, which is different than our two-step problem. Influence functions based on ordinary derivatives are used there, though in the estimation step to aid with tree splitting, rather than for inference per se, as we do below. Foster and Syrgkanis (2023) are focused on risk minimization, as opposed to inference, but consider a similar class of models and use orthogonal scores to obtain improved properties. Their estimation target more closely resembles our two step problem, though a key innovation in their case is performing this in one step, rather than estimating the primitives and then studying different counterfactual questions. Their first and second step parameters can also be more general objects than ours in some ways.

4 Inference

For second step inference on $\boldsymbol{\mu}_0 = \mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}), \boldsymbol{t}^*)]$, we apply and contribute to the recent semiparametric inference literature and the now-common double machine learning (DML) method of Chernozhukov et al. (2018a). DML is a generic method for obtaining semiparametric inference that combines two ingredients: sample splitting and a Neyman orthogonal score. Asymptotic Normality then follows under weak conditions on first step estimators, which is particularly important for ML-based first steps, because relatively little is known of their mathematical and statistical properties. Typically, L_2 convergence rates are sufficient (along with mild regularity conditions). See Chernozhukov et al. (2018a), Newey and Robins (2018), and Chernozhukov et al. (2022a) for complete discussion and further references. As usual with DML, our results apply to any first-step estimator that has fast enough rates. However, an automatic differentiation engine may be required in order to compute the influence function.

While sample splitting is straightforward both conceptually and in applications, orthogonal scores are not, and this is where our contribution lies. The key characterization of a Neyman orthogonal

score is that is has a zero derivative with respect to the first stage parameters (see Section 6), which translates to less sensitivity to first step error. Though more general conceptually, Neyman orthogonal scores often come from influence functions where this zero derivative property is ensured. Knowing the influence function, or orthogonal score, in advance is required for using these methods, and this is often a hurdle in applications. For this reason many papers stick to known examples (chiefly average treatment effects or partially linear models), list references where scores are derived, or hand-derive new scores (typical examples are Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), and Chernozhukov et al. (2022a)). We contribute to this inference method by showing that an influence function for $\boldsymbol{\mu}_0$ is automatically available for *any* combination of $\ell(\cdot)$ and $\boldsymbol{H}(\cdot)$ in Equations (3.3) and (3.4) provided their ordinary derivatives exist. We give the form of this influence function and discuss how it can be automatically computed in applications, even when it cannot be derived precisely or written down. This makes it easy to deploy in practice, because $\ell(\cdot)$ and $\boldsymbol{H}(\cdot)$ are defined by the researcher, and the rest can proceed automatically, requiring only nonparametric regression. A key insight into applicability is that the orthogonal score need not be known as a function per se, we only require that its value be computable at each observation. Our goal and results here are closest to the series of work "auto-DML", see Remark 3 below.

To state the influence function result, and thus estimation and inference for $\boldsymbol{\mu}_0$, we define relevant derivatives, which may be known or found with automatic differentiation. Let $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \cdot; \boldsymbol{t}^*)$ be the $d_{\boldsymbol{\mu}} \times d_{\boldsymbol{\theta}}$ Jacobian of $\boldsymbol{H}$ with respect to $\boldsymbol{\theta}$. Let $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \cdot)$ be the gradient of ℓ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \cdot)$ be the matrix of second derivatives. These are *ordinary* derivatives, not functional derivatives, and are thus computable automatically, even if they are evaluated at the value of the function $\boldsymbol{\theta}_0(\boldsymbol{x})$.[3] Then an influence function that applies to any combination of enriched model (3.3)

[3]To be precise, $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ is the $d_{\boldsymbol{\theta}}$-vector of first derivatives with respect to the parameter, evaluated at the number $\boldsymbol{\theta}(\boldsymbol{x})$, as in

$$\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) = \left. \frac{\partial \ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b})}{\partial \boldsymbol{b}} \right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x})}.$$

The gradient $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \cdot; \boldsymbol{t}^*)$ is similar. Then $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ is the $d_{\boldsymbol{\theta}} \times d_{\boldsymbol{\theta}}$ matrix of second derivatives, with $\{k_1, k_2\}$ element given by

$$\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))\right]_{k_1,k_2} = \left. \frac{\partial^2 \ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b})}{\partial b_{k_1} \partial b_{k_2}} \right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x})},$$

where b_{k_1} and b_{k_2} are the respective elements of the place-holder $\boldsymbol{b}$. The use of standard differentiation in these contexts has been used in some prior work, though to our knowledge not paired with automatic differentiation to obtain feasible inference in such a broad set of models. One can obtain the influence function using functional derivatives and then evaluate these at the true function $\boldsymbol{\theta}_0(\cdot)$, but our point is that one only needs the ordinary derivative evaluated at the number $\boldsymbol{\theta}_0(\boldsymbol{x})$ (or the data points more specifically).

and parameter of interest (3.4) is

$$\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*\right) - \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})), \tag{4.1}$$

where $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}]$ the population conditional Hessian of ℓ, all evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x})$ and is the "other" nonparametric object that generally arises in the correction term. The (inverse) propensity score is perhaps the most familiar example, and our requirements on $\boldsymbol{\Lambda}(\boldsymbol{x})$ mirror that case exactly. Theorem 2 below justifies this result and discusses it further.

To build intuition, it is useful to note that our result has precisely the same form as its parametric counterpart, but appropriately generalized. For the parametric two step inference problem of (3.1) and (3.2), the influence function is known to be

$$\boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}; \boldsymbol{t}^*\right) - \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}; \boldsymbol{t}^*)\boldsymbol{\Lambda}^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}),$$

where $\boldsymbol{\Lambda} = \mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta})]$ (Newey and McFadden, 1994). Equantion (4.1) is the same, and equally general, but enriched and hence conditional on $\boldsymbol{X} = \boldsymbol{x}$. This tight connection helps with implementation, because if the original structural model is understood by the researcher, so is the enriched version. For example, identification in the parametric case typically requires $\boldsymbol{\Lambda}^{-1}$ to exist, and in the enriched version this must hold for all $\boldsymbol{x}$. This is often a matter of assuming a positive conditional variance, rather than marginal variance.

Implementation of DML, once the influence function is known, has become standard, and we will not give details here to save space. Some discussion is in Appendix A.1. The key novelty here is that all the elements (4.1) can be computed automatically, if not known in advance, including the values $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}_i, \boldsymbol{t}_i, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, which can then be projected onto $\boldsymbol{x}_i$ by any standard nonparametric or machine learning method. This may require three-way sample splitting, so that $\boldsymbol{\theta}(\boldsymbol{x})$ and $\boldsymbol{\Lambda}(\boldsymbol{x})$ are estimated on separate data, which is a slight change relative to the usual DML case.

Remark 1 (Efficiency)**.** In many cases, our influence function matches the efficient one. When the original model is based on a likelihood or exponential family, and (3.3) and (3.4) contain all information, we conjecture that efficiency is always obtained following Remark 4.1 of Mammen and van de Geer (1997). In general, however, our estimator result is not guaranteed to be efficient. For

example, in partially linear models we only obtain efficiency under homoskedasticity assumptions (Appendix B). ♣

Remark 2 (Notes on $\boldsymbol{\Lambda}(\boldsymbol{x})$)**.** The function $\boldsymbol{\Lambda}(\boldsymbol{x})$ is a nuisance in the truest sense: it is required only because we use influence functions as a tool to obtain valid inference. In our case, $\boldsymbol{\Lambda}(\boldsymbol{x})$ is always low-dimensional and consists only of regressions, not conditional density functions. In some cases obtaining $\boldsymbol{\Lambda}(\boldsymbol{x})$ is simplified. If $\boldsymbol{T}$ is randomly assigned, or more generally independent of $\boldsymbol{X}$, then $\boldsymbol{\Lambda}(\boldsymbol{x})$ can often be computed or estimated more simply, though it may remain a function of $\boldsymbol{x}$. Even if not randomly assigned, if $\boldsymbol{T}$ is known to be assigned based on a subvector of $\boldsymbol{X}$, such as in targeting problems, this can be imposed on the estimation. Inverse functions are standard in semiparametric inference, and in practice this piece is often the most difficult. The challenge is generally model- and data- specific, not related to the choice of nonparametric/ML method. Often some form of regularization is used. The most widely known is trimming the propensity score, which has received rigorous study by Ma and Wang (2020). Extending this work to our setting would be useful. ♣

Remark 3 (Auto-DML)**.** The recent work on "auto-DML" shares our goal of applying DML without having to derive a new orthogonal score each time. This line of work (Chernozhukov, Newey, and Singh, 2022b,a; Chernozhukov et al., 2022a,b, 2024a,b) shows that the correction term of the influence function satisfies certain moment conditions, and these can be used to form estimates. In the context of (4.1), this amounts to estimating the entire quantity $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}$. There are different versions of this method, but overall, it is as general as our approach. Our approach can be applied to many second stage parameters, since the first step correction need only be estimated once. We do require that $\boldsymbol{\theta}_0(\boldsymbol{x})$ enter the second stage only through its evaluation at a data point. The auto-DML method estimates the inverse directly, which can yield more stable performance than estimating $\boldsymbol{\Lambda}(\boldsymbol{x})$ then inverting. On the other hand, examining $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is often a key step in the analysis, such as examining estimated propensity scores to evaluate the overlap assumption. ♣

Remark 4. Influence functions have appeared in many different contexts in statistics recently and our results can potentially be used to extend these methods to new contexts. To give several examples:

- Our main use is for DML, which is for debiasing inference following Chernozhukov et al. (2018a) and Newey (1994) more generally. Similar debiasing using correction terms appears in high dimensional methods (Belloni, Chernozhukov, and Hansen, 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014) and for loss functions directly (Foster and Syrgkanis, 2023).

- Athey and Wager (2021) study policy optimization and show the remarkable finding that using an orthogonal score yields faster reminder rates in terms of welfare just as for inference. Our score could be used to bring their insight into other areas.

- Koh and Liang (2017) use influence functions to try to understand "black-box" ML methods. Extending this to economic contexts would be valuable in applied research and policy evaluation.

- Robins et al. (2008) use higher-order influence functions to obtain refined semiparametric inference. We conjecture that using automatic differentiation could be used to obtain higher order inference just as with our first order results.

In addition, some approaches in Remark 6 employ orthogonal scores, though for estimation. ♣

Remark 5 (Series Approach)**.** Intuitively, our derivation treats the functions $\boldsymbol{\theta}_0(\boldsymbol{x})$ themselves as a vector of parameters, and we obtain a result familiar from parametric two-step models (Newey and McFadden, 1994). An alternative approach is to consider $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ as a parametric model (where the weights and biases of the deep net are the parameters) and apply the same two-step estimation logic. This is shown to be a valid approximation to the semiparametric case in some contexts by Ackerberg, Chen, and Hahn (2012). Applying this idea to deep learning may be valid, but is practically infeasible as the number of parameters is too large and the estimator too complex. For example, the equivalent of $\boldsymbol{\Lambda}(\boldsymbol{x})$ would be a square matrix of dimension equal to the number of parameters in the deep net, which can be extremely large. Computing and inverting such a matrix may be impossible. ♣

5 Empirical Applications

5.1 Application: Advertising and Personalized Interest Rates

5.1.1 Empirical Context

In this section we use our framework to replicate and extend the analysis presented in Bertrand et al. (2010). The data is from a large scale field experiment run on behalf of a financial institution in South Africa. Consumers were sent marketing material for short terms loans where a number of features of the advertising content and the interest rate offered were all randomized (full details are left to that paper). The vector of treatments is thus $\boldsymbol{T} = (\boldsymbol{C}', R)'$, where $\boldsymbol{C}$ denotes the advertising content and R the interest rate offered. The key outcome variable (Y) is the indicator for whether or not the consumer applied for the loan. We will use a binary choice model, one of the workhorse models in applied economics. Other variables following the application were also tracked such as a default indicator (D) and the loan amount (L). The data also contains a rich set of demographics ($\boldsymbol{X}$) which we use to calibrate our measures of heterogeneity.

We conduct two analyses using their data. First, we replicate their analysis and extend it to allow for heterogeneity captured by our structured DNNs. We use our specification to compute the average marginal effect of each treatment and compare those to the results presented by the authors. Second, we use the results of the model (with some additional assumptions) to construct optimal personalized interest rate offers and compute the expected profits from implementing the personalization scheme, and we then conduct inference using our novel methodology.

5.1.2 Model and Implementation

We enrich the setup in Bertrand et al. (2010) and assume that consumers have a utility

$$u_i = \boldsymbol{\theta}_{\boldsymbol{C}}(\boldsymbol{x}_i)'\boldsymbol{c}_i + \theta_R(\boldsymbol{x}_i) r_i + \varepsilon_i,$$

where the vector of parameter functions $\boldsymbol{\theta}(\boldsymbol{x}) = (\boldsymbol{\theta}_{\boldsymbol{C}}', \theta_R)'$ is partitioned to match $\boldsymbol{T} = (\boldsymbol{C}', R)'$. We assume that the ε_i are i.i.d. Logistic, yielding the standard probabilities of response:

$$\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G\left(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}\right) = \frac{1}{1 + \exp\left(-\left[\boldsymbol{\theta}_{\boldsymbol{C}}(\boldsymbol{x})'\boldsymbol{c} + \theta_R(\boldsymbol{x}) r\right]\right)}.$$

Using these probabilities we can construct the log-likelihood as

$$\ell(y, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) = y \log \left(\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] \right) + (1 - y) \log \left(\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] \right), \quad (5.1)$$

which has been enriched from the standard version. The negative of this will serve as the loss (**??**) for our problem. One can easily verify the high-level assumptions in this setting, particularly given that the binary choice model is widely studied and well understood. For example, it is straightforward to find that $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[G\left(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}\right)\left(1 - G\left(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}\right)\right)\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X} = \boldsymbol{x}]$, which will be invertible under standard and commonly used economic assumptions.

We will obtain parameter function estimates by solving (**??**) using this $\ell(y, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ and the architecture Figure **??**. In our implementation we approximate the $\boldsymbol{\theta}(\boldsymbol{x})$ via a simple deep neural networks with two hidden layers with 80 and 40 nodes, respectively. Part of the simplicity of the network architecture stems from the fact that we have a smallish dataset ($N = 53194$) and rather large dimension of the treatment vector ($d = 13$). For inference purposes with used three-fold cross fitting, using two thirds of the data to obtain $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ and one third to obtain $\widehat{\boldsymbol{\mu}}$.

5.1.3 Results: Marginal Effects

The parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ are the key inputs into the target of inferential interest, $\boldsymbol{\mu}_0$. We will use our estimated $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ and novel influence function to examine the marginal effect of the treatments and compare them to those presented in column (1) Table III of Bertrand et al. (2010) (on pages 291-294). We note that our specifications are slightly different since they use a Probit specification while we use the Logit. Even so, our approach will allows us to comment on the heterogeneity in said effects.

Binary choice models are widely used in applications, often with price as (at least one of) the treatment variable(s). Our framework would immediately give inference for many standard, policy-relevant parameters in this context. Examples, beyond those shown below, include (i) the price elasticity at a price (here, interest rate) r, which sets $H = (1 - \mathbb{P}[Y = 1|\boldsymbol{x}, r])\theta_R(\boldsymbol{x})r$; (ii) a measure of willingness to pay obtained by taking $H = \theta_R(\boldsymbol{x})/\theta_{01}(\boldsymbol{x})$, where θ_{01} is the intercept; and (iii) expected consumer welfare, $H = -\theta_R(\boldsymbol{x})^{-1} \log(1 + \exp(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}))$. Importantly, without our explicit use of an enriched structural model, it would not be easy to characterize these parameters

Table 1: Results

Variable	DNN-AME	95%CI(L)	95%CI(U)	QJE	$\Pr(\theta(x) > 0)$	Coef. of Var.
Interest Rate Offer	-0.0047	-0.0083	-0.0011	-0.0029	0.1337	1.0211
We speak your language	-0.0048	-0.0137	0.0041	-0.0043	0.2533	2.0542
Special rate for you	-0.0034	-0.0120	0.0053	0.0001	0.5001	4.4506
No photo	0.0038	-0.0060	0.0136	0.0013	0.5723	3.4931
Black photo	0.0016	-0.0064	0.0096	0.0058	0.5402	5.1348
Female photo	0.0060	-0.0021	0.0141	0.0057	0.6820	2.3375
Cell phone raffle	-0.0009	-0.0104	0.0085	-0.0023	0.4812	17.0059
Example loan shown	0.0044	-0.0084	0.0173	0.0068	0.8631	1.9379
No loan use mentioned	0.0108	0.0009	0.0207	0.0059	0.7499	1.0936
Interest rate shown	0.0017	-0.0085	0.0119	0.0025	0.6289	7.9903
Loss comparison	0.0001	-0.0081	0.0083	-0.0024	0.2342	89.3606
Competitor rate shown	0.0013	-0.0085	0.0111	-0.0002	0.4107	9.6790

and obtain inference.

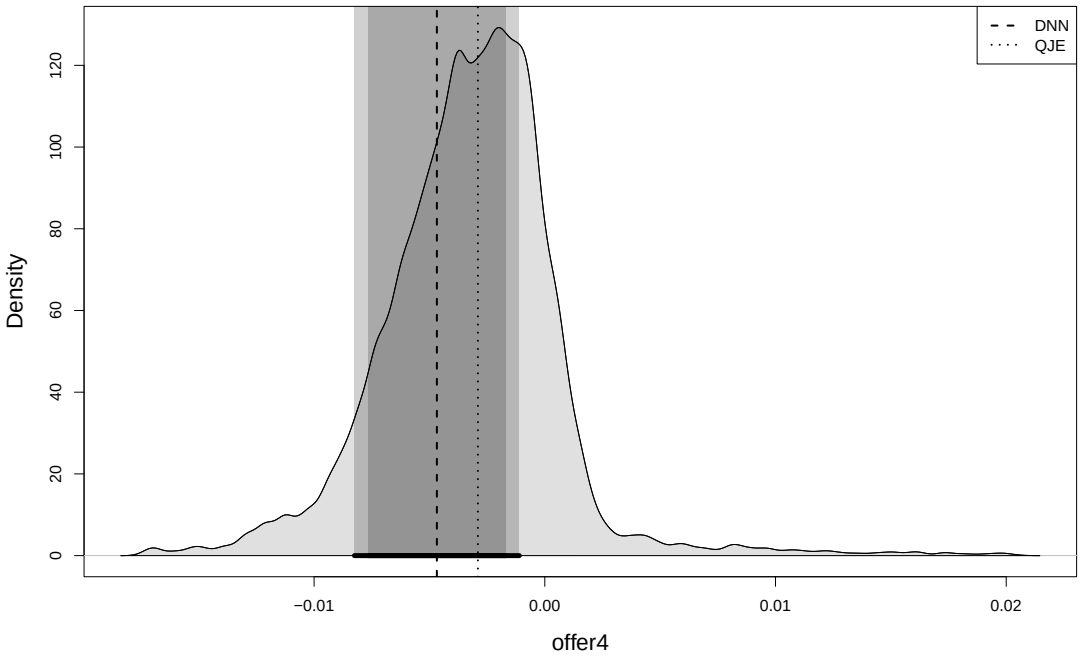
For our specification, the average marginal effects for any given treatment can be written in closed form. Recall that $\boldsymbol{\theta}(\boldsymbol{x}) = (\boldsymbol{\theta}'_{\boldsymbol{C}}, \theta_R)'$ is partitioned to match $\boldsymbol{T} = (\boldsymbol{C}', R)'$. Then, for example, the average marginal effect of a change in interest rates is

$$\text{AME}(R) = \mathbb{E}\left[\frac{\partial G(\boldsymbol{\theta}_0(\boldsymbol{X})'\boldsymbol{t})}{\partial r}\bigg|_{\boldsymbol{t}=\boldsymbol{t}^*}\right] = \mathbb{E}\left[G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}^*)\left(1 - G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}^*)\right)\theta_R\right].$$

In our empirical results we set $\boldsymbol{t}^*$ to the sample average for simplicity. Similarly quantities for the advertising content are readily available. Thus, by taking $\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*) = G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}^*)(1 - G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}^*))\theta_R$ we can apply our framework, obtaining inference post-DNN easily.

We present the results of our analysis in Table 1. For convenience we have also included the corresponding estimates of Bertrand et al. (2010) under the column heading **QJE**. As is evident, the deep net estimates match up with the results of the original paper quite well with the confidence interval containing the original estimates, which we can interpret as the original, overall, findings being robust to heterogeneity. In addition, we also uncover substantial heterogeneity in the estimates as demonstrated by the coefficient of variation of $\boldsymbol{\theta}(\boldsymbol{x})$. A more useful depiction of these results are contained in Figures 8 and 9 which showcase the rich heterogeneity in the estimated effects. In each we present 90% and 95% confidence intervals for the average marginal effect, with vertical lines to indicate the estimate from our analysis as well as those from Bertrand et al. (2010). We see that although the average matches up roughly with the rigid parametric model, substantial heterogeneity exists, which will be important for targeting. As discussed, other statistics of interest can be easily computed following a similar procedure.

Figure 8: Marginal Effect of Interest Rate



5.2 Discrete Choice Models

Our second application uses the context of discrete choice models with panel data to showcase our proposed framework. Discrete choice models (McFadden and Train (2000)) are ubiquitous across the social sciences. The basic framework aims to model agent choices as a function of treatments (such as prices) and possibly covariates. In this setting, agents make choices by maximizing some latent payoff function that reflects both observable attributes of each option and unobservable idiosyncrasies. In economics, discrete choice models are used in a variety of applications including labor market decisions, such as the selection between employment offers (e.g., private vs. public sector jobs), health economics, where they are used to analyze patient and provider choices, such as the selection of healthcare plans or treatment options and in Industrial organization and marketing where product or brand choices are made as a function of prices and advertising.

In what follows we will showcase our framework and its application to heterogeneous discrete choice models. Heterogeneity in discrete choice models is typically accounted for by allowing variation in the parameters governing agents' preferences and sensitivities to treatment effects. The most

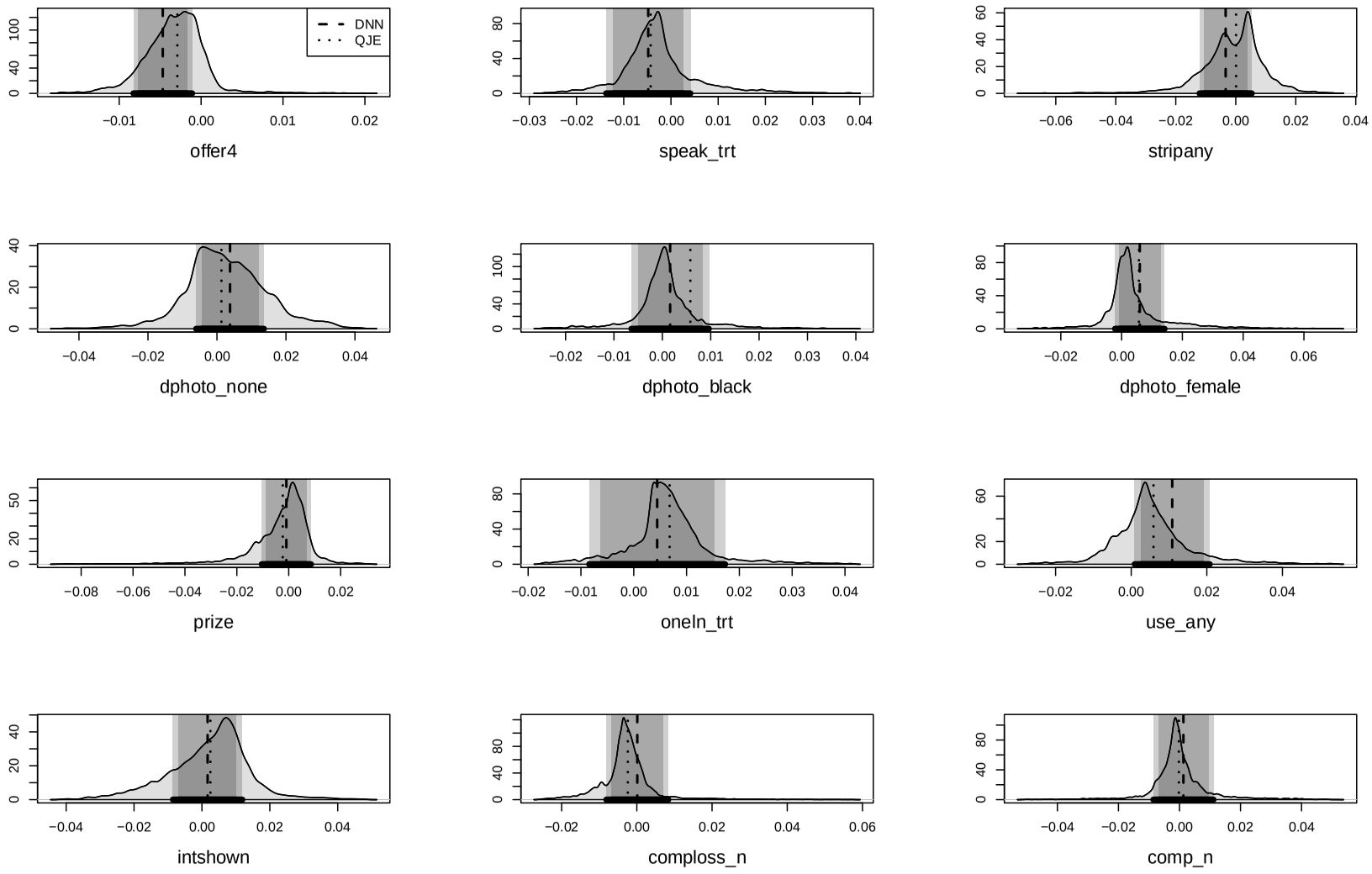


Figure 9: Marginal Effect of Advertising Content

common methods for estimating this heterogeneity include random coefficient models (often termed mixed logit models), where parameters are treated as random variables drawn from a specified distribution which may be continuous or discrete. There are also approaches that treat heterogeneity via some form of individual or group fixed effects. Our approach takes a slightly different approach by projecting the heterogeneity onto observed individual characteristics as described earlier in this paper.

Our application context closely follows a recent paper by Smith, Seiler, and Aggarwal (2022), who utilize household scanner panel data provided by Information Resources, Inc. (IRI) to model consumer brand choice in mayonnaise category. The data comes from a single market (Pittsfield, Massachusetts) and was collected over a 22-month period from January 2011 to October 2012. The dataset includes information from four stores belonging to two retail chains, representing consumer purchasing behavior. As is usual, the data contain weekly pricing and promotional information for three mayonnaise brands: Hellmann's, Kraft Miracle Whip, and Cain's. Elements such as retail prices, discounts, feature advertising, and product displays are included which allow us to analyze the impact of marketing activities on consumer purchasing decisions.

The data describes choices from a panel of 581 households each of which had at least three purchases in the mayonnaise category and consistent store visits throughout the study period. The data spans 95,949 shopping trips, resulting in 3,394 purchases. Finally, and relevant to our framework, the data includes basic demographic information, including household income, family size, and employment status, as well as characteristics such as age and education. For further details on the data we refer the reader to the original study by Smith, Seiler, and Aggarwal (2022).

5.2.1 The Choice Framework

Consider the following utility specification,

$$U_{ijs} = \alpha_j(x_i) + \beta(x_i)'\boldsymbol{t}_{js} + \epsilon_{ijs}.$$

The above reflects the utility consumer i derives from purchasing brand j sin shopping trip s. $\alpha_j(x_i)$ reflect consumer i's baseline preference for brand j while the $\beta(x_i)$ connote the sensitivity to marketing elements $\boldsymbol{t}$. In the application the vector $\boldsymbol{t}$ will contain brand prices and other marketing

activities such as feature advertising and display. As before we will collect parameter functions and say $\theta(x_i) = \{\{\alpha_j(x_i)\}_{j=1}^J, \beta(x_i)\}$. Note that, as before, we are assuming that $\theta(x_i)$ approximates some θ_i.

The fact that the context contains a panel data does not require any adjustments to our framework. To see this first note that

$$P(Y_{ijs} = 1|x_i, \boldsymbol{t}) = \frac{\exp(\alpha_j(x_i) + \beta(x_i)'\boldsymbol{t}_{js})}{1 + \sum_{l=1}^J \exp(\alpha_l(x_i) + \beta(x_i)'\boldsymbol{t}_{ls})}$$

and write the sample log-likelihood as

$$\ell(\boldsymbol{Y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) = \sum_{i=1}^N \sum_{t=1}^{T_i} Y_{ijs} \ln P(Y_{ijs} = 1|x_i, \boldsymbol{t}) \tag{5.2}$$

In the above we can think of $\boldsymbol{Y}$ as a vector valued outcome variable and as such $\ell(\boldsymbol{Y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ is simply some loss function that adheres to our requirements for analysis and inference. More generally, we could also allow for random coefficients or finite mixtures since those do not alter the conditions we require for our framework to work.

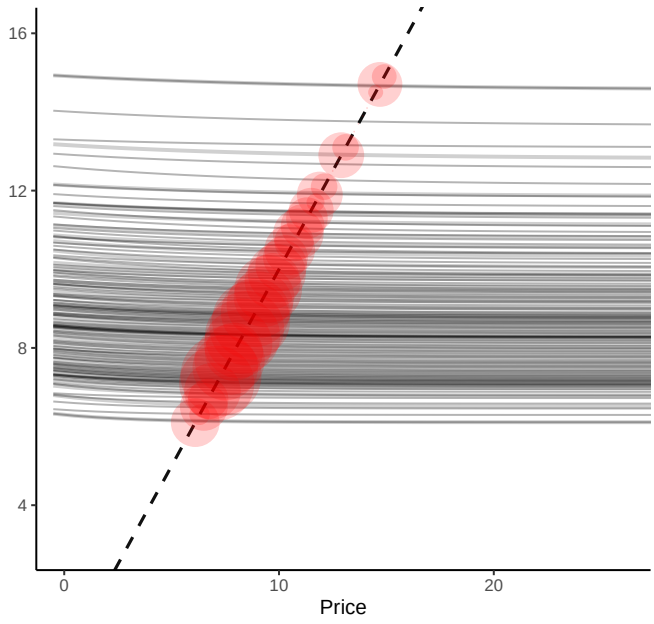
5.2.2 Implementation

We tested a number of architectures when estimating (5.2). In particular, we examines between one and three layers with 32 or 64 nodes each. We also explored various regularization approaches. Results were reasonably stable across architectures. We used 10-fold cross-fitting with 5 folds for estimation of $\boldsymbol{\theta}(\boldsymbol{x})$, 4 folds for estimation of Λ and the last fold for averaging. We note here that out cross-fitting is cluster conforming. Finally we used median aggregation over 5 cross-fitting repetitions. Our estimator for Λ used a "short-stacking" approach with an mixture across various ML algorithms including linear regression, lasso and variants, random forests and boosting trees. Estimates of Λ^{-1} were obtained via Moore-Penrose pseudo inverses with the shrinkage parameter set to 1e-5.

5.2.3 Results: Inference for Counterfactual Profits

The directions and magnitude for the parameter functions estimated conformed to expectations and we do not belabor their discussion here to conserve space. Instead we focus on demonstrating

Figure 10: Optimal Prices



how our approach can be used to construct valid confidence intervals for equilibrium counterfactual objects of interest.

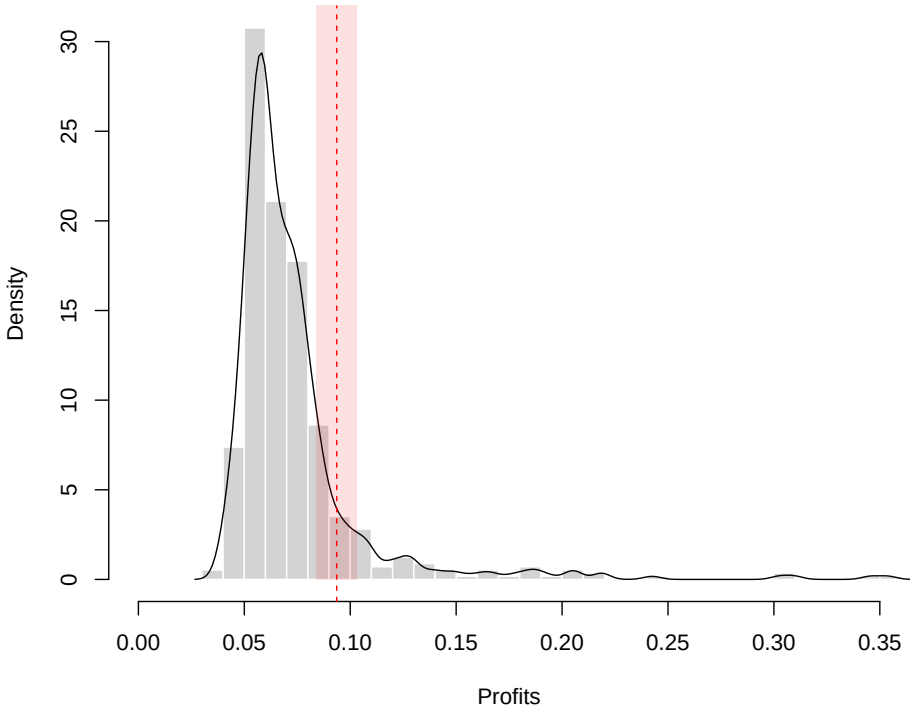
We start by using the estimates to obtain profit maximizing prices $p^*(x_i)$ which are based on maximizing expected profits

$$p_j^*(x_i) = \arg\max_p P(Y_{ijs} = 1|x_i, p, \tilde{\boldsymbol{t}})(p - mc)$$

where we focus on a single brand and fix all other marketing elements (as well as competitor prices) to some predetermined level. Marginal costs were set to 2.49 as in the original paper. The distribution of optimal prices is depicted in (10). We then use these prices in conjunction with our framework to construct confidence intervals for expected profits under these optimal prices. These are depiced in (11).

While we acknowledge that the specification, the assumptions around constructions of optimal prices and profits might be somewhat arbitrary, they highlight the flexibility and ease of using our framework for inference around sophisticated economic constructs. We could easily use the same approach to compute elasticities, welfare measures or to test hypotheses related to competing counterfactual assumptions.

Figure 11: Profits at Optimal Prices



6 Theoretical Results

In this section we state the formal theoretical results that underpin the methodology of Sections 3 and 4. We establish (i) nonasymptotic error bounds (and implied convergence rates) for our novel structured deep learning, (ii) the validity of the influence function as an orthogonal score, and (iii) asymptotic normality of the second step estimator. All proofs are given in Appendix C.

Our deep learning theory generalizes Farrell, Liang, and Misra (2021) to the structural setting. We impose two assumptions. For the loss function, we require Lipschitz continuity in general and, near the truth, sufficient curvature. Neither are restrictive and both are common in the nonparametric M estimation literature (cf Chen (2007) and others, where further references and use of other norms are discussed). These conditions are for estimation of $\boldsymbol{\theta}_0(\boldsymbol{x})$; further assumptions will be required for inference.

Assumption 1. *Suppose that $\boldsymbol{\theta}_0(\boldsymbol{x})$ are nonparametrically identified in (3.3), uniformly bounded, and that there are constants c_1, c_2, and C_ℓ that are bounded and bounded away from zero, such that $|\ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}(\boldsymbol{x})) - \ell(\boldsymbol{y},\boldsymbol{t},\tilde{\boldsymbol{\theta}}(\boldsymbol{x})| \leq C_\ell \|\boldsymbol{\theta}(\boldsymbol{x}) - \tilde{\boldsymbol{\theta}}(\boldsymbol{x})\|_2$ and*

$$c_1 \mathbb{E}\left[\|\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] \leq \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))] - \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))] \leq c_2 \mathbb{E}\left[\|\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right].$$

These requirements will often be implied by restrictions on the gradient and Hession of the loss, or on the matrix $\boldsymbol{\Lambda}(\boldsymbol{x})$. Such restrictions are natural in our setting, since they are applied to parametric structural models and differentiability will be required for inference later on. However, differentiability is not required here and thus our results can be used in nonsmooth cases (for example Tambwekar et al. (2022) and Padilla, Tansey, and Chen (2022) apply the theory of Farrell, Liang, and Misra (2021) to quantile regression and the same extension could be done here).

Let $\boldsymbol{W} = (\boldsymbol{Y}', \boldsymbol{T}', \boldsymbol{X}')'$ be the population random variables, with an observation denoted $\boldsymbol{w}_i = (\boldsymbol{y}_i', \boldsymbol{t}_i', \boldsymbol{x}_i')'$. Let $\boldsymbol{X}_c$ denote the continuously distributed elements of $\boldsymbol{X}$, and define $d_c = \dim(\boldsymbol{X}_c)$, and take the rest to be binary random variables, without loss of generality. Part (iii) of this assumption restricts to smooth functions, which are known to be approximable by deep neural networks (Yarotsky, 2017, 2018; Hanin, 2017).

Assumption 2. *(i) the elements of $\boldsymbol{W}$ are bounded random variables. (ii) $\boldsymbol{X}_c$ has compact connected*

support, taken to be $[-1,1]^{d_c}$. *(iii) As functions of* $\boldsymbol{x}_c$, *the continuously distributed components of* $\boldsymbol{X}$, $\theta_{0k}(\boldsymbol{x}) \in \mathcal{W}^{p,\infty}([-1,1]^{d_c})$, *for* $k = 1, \ldots, d_{\boldsymbol{\theta}}$, *where for positive integers* p *and* q, *define the Hölder ball* $\mathcal{W}^{p,\infty}([-1,1]^q)$ *of functions* $h : \mathbb{R}^q \to \mathbb{R}$ *with smoothness* $p \in \mathbb{N}_+$ *as*

$$\mathcal{W}^{p,\infty}([-1,1]^q) := \left\{ h : \max_{\boldsymbol{r}, |\boldsymbol{r}| \leq p} \operatorname*{ess\,sup}_{\boldsymbol{v} \in [-1,1]^q} |D^{\boldsymbol{r}} h(\boldsymbol{v})| \leq 1 \right\},$$

where $\boldsymbol{r} = (r_1, \ldots, r_q)$, $|\boldsymbol{r}| = r_1 + \ldots + r_q$ *and* $D^{\boldsymbol{r}} h$ *is the weak derivative. (iv)* $\max_{k \leq d_{\boldsymbol{\theta}}} \sup_{\boldsymbol{x}} |\theta_{0k}(\boldsymbol{x})| < M$, *for some positive* M.

Our result for deep learning is the following. We specialize to the standard implementation using deep and wide multi-layer perceptrons (fully connected, feedforward neural networks). Equation (C.4) in Appendix C shows a more general bound that is agnostic about the type of approximation, and hence the type of network. This can be used to obtain faster rates or cover fixed-width, very deep networks (see Section 2.3 of Farrell, Liang, and Misra (2021) for discussion).

Theorem 1. *Let* $\boldsymbol{w}_i$, $i = 1, \ldots, n$, *be a random sample that obeys Assumptions 1 and 2. Define* $\widehat{\boldsymbol{\theta}}$ *as the estimator found by solving* (3.5), *where the class* $\mathcal{F}_{DNN}$ *is a feedforward, fully connected network with ReLU activation structured according to Figure 7, without parameter functions by* $2M$. *If the width and depth are such that* $H \asymp n^{(d_c)/2(p+d_c)} \log^2 n$ *and* $L \asymp \log n$, *respectively, then*

$$\|\widehat{\theta}_k - \theta_{0k}\|^2_{L_2(\boldsymbol{X})} \leq C \cdot \left\{ n^{-\frac{p}{p+d_c}} \log^8 n + \frac{\log \log n}{n} \right\}$$

and

$$\mathbb{E}_n \left[\left(\widehat{\theta}_k - \theta_{0k} \right)^2 \right] \leq C \cdot \left\{ n^{-\frac{p}{p+d_c}} \log^8 n + \frac{\log \log n}{n} \right\},$$

for n *large enough, with probability* $1 - \exp\{-n^{\frac{d_c}{p+d_c}} \log^8 n\}$, *for* $k = 1, \ldots, d_{\boldsymbol{\theta}}$, *where the constant* C *may depend on the* $\dim(\boldsymbol{W})$, $d_{\boldsymbol{\theta}}$, *and other fixed quantities in Assumptions 1 and 2.*

The result of Theorem 1 speaks directly to the nonparametric M estimation literature. This theorem takes deep learning away from prediction and toward learning economically meaningful parameters. It shows that deep nets enjoy the same properties of other nonparametric/ML methods, but with the advantages of structure. A theoretical drawback is that for a given smoothness level, this rate is not optimal. Obtaining the optimal rate, or studying different norms, would be a

useful extension. The bounds are sharp enough for inference and reflects the excellent empirical performance.

Remark 6 (CATE Estimation)**.** Our results in this context are loosely connected to the large recent literature on nonparametric estimation of the conditional average treatment effect (CATE) function for a binary treatment, which corresponds to $\boldsymbol{\beta}_0(\boldsymbol{x})$ in (B.2) with identity $G(u)$ and a scalar, binary $\boldsymbol{T}$. Here $\beta(\boldsymbol{x})$ is the CATE and can be written as the different of two prediction functions. Many results, including using neural networks, show that under certain conditions $\beta(\boldsymbol{x})$ can be estimated faster than the typical nonparametric bound of Stone (1982), which is itself faster than Corollary B.1. Generally, this is possible when $\beta(\boldsymbol{x})$ is somehow "simpler" than its components, such as being smoother or lower dimensional. See, for example, Shalit, Johansson, and Sontag (2017) and Nie and Wager (2020) for early important work Kennedy (2023) for fast rates and much other discussion, Foster and Syrgkanis (2023) for global bounds including sharp risk bounds, and finally Kennedy et al. (2024) for quantifying the minimax bound precisely. We do not seek any form of adaptivity here (see also Remark B.2) but it is an interesting conjecture that the same performance for CATE estimation carries over to other economic contexts, particularly since some of these papers apply orthogonal scores. ♣

Next we give our influence function result. Influence functions have a long history in econometrics. Newey (1994) remains the seminal treatment. We defer to that work and Ichimura and Newey (2022) for background and theory of influence functions, including regularity conditions for their existence. Our goal is not to contribute to the theory of influence functions per se, but rather to use the tools of this theory to obtain the methodological result of a broadly applicable influence function, to enable two-step semiparametric inference under weak conditions. That is, we view the influence function as a tool for obtaining feasible inference, rather than an object of interest in its own right. This viewpoint is implicit in recent work on inference after ML (e.g., Belloni, Chernozhukov, and Hansen, 2014; Farrell, 2015; Chernozhukov et al., 2018a) but it is worthwhile to make it explicit to better understand how this mode of thinking allows us to cover such a wide range of applications and the apply automatic differentiation.

The assumption we impose next is mostly standard and ensures sufficient regularity for our influence function to be calculated and for asymptotic Normality of the resulting estimator. One

conceptual point is that further assumptions will be needed for a causal interpretation, such as unconfoundedness or conditional exogeneity.

Assumption 3. *The following conditions hold on the distribution of* $\boldsymbol{W} = (\boldsymbol{Y}', \boldsymbol{X}', \boldsymbol{T}')'$*, uniformly in the given conditioning elements. (i) Equation* (3.3) *holds and identifies* $\boldsymbol{\theta}_0(\boldsymbol{x})$*, where* $\ell(\boldsymbol{w}, \boldsymbol{\theta})$ *is thrice continuously differentiable with respect to* $\boldsymbol{\theta}$*. (ii)* $\mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = 0$*. (iii)* $\boldsymbol{\Lambda}(\boldsymbol{x})$ *is invertible with bounded inverse. (iv) The parameter* $\boldsymbol{\mu}_0$ *of Equation* (3.4) *is identified and pathwise differentiable and* $\boldsymbol{H}$ *is thrice continuously differentiable in* $\boldsymbol{\theta}$*. (v)* $\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)$ *and* $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X}))$ *possess* $q > 4$ *finite absolute moments and positive variances.*

The most important assumptions here are that the first order condition of (3.3) holds and that $\boldsymbol{\mu}_0$ is pathwise differentiable. The latter keeps focus on regular semiparametric contexts. The former follows our idea to take a well-defined parametric model, for which such identification would hold, and enrich the model with machine learning. Conditional mean restrictions are a particularly popular case. Condition (iii) will often be implied by other conditions on the model, such as in the case of logistic regression if $\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}]$ is bounded away from zero and one (which in turn may be implied by conditions on $\boldsymbol{X}$, $\boldsymbol{T}$, and the functions $\boldsymbol{\theta}_0$, such as boundedness). Or, in the context of treatment effects we need the standard overlap condition. Some version of the condition of positive variance, or invertibility of $\boldsymbol{\Lambda}(\boldsymbol{x})$, is quite standard in semiparametric problems.

The following result justifies our inference method. Other than the novel influence function, it is an application of existing DML theory (Chernozhukov et al., 2018a). Let $\mathbf{0}_d$ be the d-long zero vector and $\boldsymbol{I}_d$ be the d-square identity matrix.

Theorem 2. *Let* $\boldsymbol{w}_i$*,* $i = 1, \ldots, n$*, be a random sample that obeys Assumption 3. Assume* $\|\widehat{\theta}_{k_1} - \theta_{0k_1}\|_{L_2(\boldsymbol{X})} = o_P(n^{-1/4})$ *and* $\|[\widehat{\boldsymbol{\Lambda}}]_{k_1,k_2} - [\boldsymbol{\Lambda}]_{k_1,k_2}\|_{L_2(\boldsymbol{X})} = o_P(n^{-1/4})$ *for all* $k_1, k_2 \in \{1, \ldots, d_{\boldsymbol{\theta}}\}$*, and that* $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{x}_i)$ *is uniformly invertible. Then (i)* (4.1) *gives a Neyman orthogonal score and (ii) DML-based* $\widehat{\boldsymbol{\mu}}$ *and* $\widehat{\boldsymbol{\Psi}}$ *obey*

$$\sqrt{|\mathcal{C}|}\widehat{\boldsymbol{\Psi}}^{-1/2}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) = \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} \boldsymbol{\Psi}^{-1/2} \boldsymbol{\psi}(\boldsymbol{w}_i, \boldsymbol{\theta}_0(\boldsymbol{x}_i), \boldsymbol{\Lambda}(\boldsymbol{x}_i))/\sqrt{n} + o_p(1) \to_d \mathcal{N}(\mathbf{0}_{d_{\boldsymbol{\mu}}}, \boldsymbol{I}_{d_{\boldsymbol{\mu}}}).$$

Here we impose the standard rate conditions on the first step estimators, allowing for valid inference after using any sufficiently accurate nonparametric/ML method. Deep learning remains

a preferred choice in structural modeling, as discussed above. The rate condition on $\boldsymbol{\Lambda}(\boldsymbol{x})$ can be weakened as shown by Chernozhukov et al. (2022a). Constructing DML-based $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Psi}}$ is standard, see Chernozhukov et al. (2018a) and Appendix A.1. For higher order asymptotic theory of DML see Velez (2024).

Theorems 1 and 2 validate our structural deep learning and inference methodology. Our assumptions are necessarily high level, given the generality of the setting. Verifying these conditions in a given setting is required, but is often straightfoward. Importantly, because we are specifically focusing on enriching a parametric structural model, familiar analyses from the parametric case can aid in interpreting the conditions required. For example, our conditions are the natural analogues of what is required in well-understood QMLE problems, and the intuition from these can be ported directly. The assumptions used here are not minimal, both for the first and second steps, and can be weakened in specific cases and with computationally intensive procedures (for recent examples, see Cattaneo and Jansson, 2018; Cattaneo, Jansson, and Ma, 2019).

Finally, it is worth noting that, like most approaches to semiparametric inference, the distributional approximation is first-order invariant to the first step. In some cases this yields a poor finite sample approximation, because different estimators will yield different results and can be sensitive to tuning parameter choices (and other implementation details) in ways not reflected in the first order distribution. Obtaining more refined results yields a superior approximation, but is only typically available in specific contexts, such as first-step kernel estimators, that allow for detailed calculation of leading terms Cattaneo, Crump, and Jansson (2014); Cattaneo, Jansson, and Newey (2018); Cattaneo et al. (2024b).

7 Conclusion

Structural modeling is a workhorse of empirical economic research. We have shown how to enirch these models with deep learning to capture rich heterogeneity, filling in the gaps left by economic theory. This method combines the strength of structural modeling and the strength of machine learning. We established convergence rates for structured deep learning and valid inference using a novel influence function calculation. Our method represents a step toward easier and more rigorous use of machine learning in economic research, but is far from complete. Shape restrictions are on

major form of structure arising from economic theory (see Chetverikov, Santos, and Shaikh (2018) for a recent review). Extending our methods to include impose shape constraints is an important step for future research. From an implementation point of view, there is also a lot of ground to cover for deep neural networks, including tuning parameter choices and robust computation.

8 References

Ackerberg, Daniel, Xiaohong Chen, and Jinyong Hahn. 2012. "A practical asymptotic variance estimator for two-step semiparametric estimators." *Review of Economics and Statistics* 94 (2):481–498. (Cited on page 18.)

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Harvard University Press. (Cited on pages 51 and 52.)

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics* 47 (2):1148–1178. (Cited on pages 4, 13, and 14.)

Athey, Susan and Stefan Wager. 2021. "Policy learning with observational data." *Econometrica* 89 (1):133–161. (Cited on page 18.)

Bach, Francis. 2017. "Breaking the curse of dimensionality with convex neural networks." *The Journal of Machine Learning Research* 18 (1):629–681. (Cited on page 46.)

Bajari, Patrick, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. 2015. "Machine learning methods for demand estimation." *American Economic Review: Papers & Proceedings* 105 (5):481–485. (Cited on page 4.)

Bartlett, Peter L, Olivier Bousquet, and Shahar Mendelson. 2005. "Local rademacher complexities." *The Annals of Statistics* 33 (4):1497–1537. (Cited on page 55.)

Bartlett, Peter L., Nick Harvey, Christopher Liaw, and Abbas Mehrabian. 2017. "Nearly-tight VC-dimension bounds for piecewise linear neural networks." In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2017)*. (Cited on page 56.)

Bauer, Benedikt and Michael Kohler. 2019. "On deep learning as a remedy for the curse of dimensionality in nonparametric regression." *Annals of Statistics* 47 (4):2261–2285. (Cited on page 46.)

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls." *Review of Economic Studies* 81:608–650. (Cited on pages 15, 18, 30, and 49.)

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei. 2016. "Post-selection inference for generalized linear models with many controls." *Journal of Business & Economic Statistics* 34 (4):606–619. (Cited on page 49.)

Berry, Steven T. 1994. "Estimating discrete-choice models of product differentiation." *The RAND Journal of Economics* 25 (2):242–262. (Cited on page 49.)

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's advertising content worth? Evidence from a consumer credit marketing field experiment." *The Quarterly Journal of Economics* 125 (1):263–306. (Cited on pages 3, 5, 6, 7, 8, 19, 20, and 21.)

Carroll, Raymond J, Jianqing Fan, Irene Gijbels, and Matt P Wand. 1997. "Generalized partially linear single-index models." *Journal of the American Statistical Association* 92 (438):477–489. (Cited on page 49.)

Cattaneo, Matias D. 2010. "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability." *Journal of Econometrics* 155 (2):138–154. (Cited on pages 4, 47, and 48.)

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng. 2023. "Nonlinear Binscatter Methods." working paper . (Cited on pages 7 and 8.)

———. 2024a. "On Binscatter." *American Economic Review* 114 (5):1488–1514. (Cited on page 49.)

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Ernst Schaumburg. 2020. "Characteristic-Sorted Portfolios: Estimation and Inference." *Review of Economics and Statistics* 101 (3):531–551. (Cited on page 49.)

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson. 2014. "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives." *Econometric Theory* 30:176–200. (Cited on page 32.)

Cattaneo, Matias D. and Max H. Farrell. 2011. "Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates." In *Advances in Econometrics: Missing Data Methods*, vol. 27A, edited by David Drukker. Emerald Group Publishing Limited, 93–127. (Cited on page 48.)

———. 2013. "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators." *Journal of Econometrics* 174:127–143. (Cited on page 13.)

Cattaneo, Matias D., Max H. Farrell, and Yingjie Feng. 2020. "Large Sample Properties of Partitioning-based Series Estimators." *Annals of Statistics* 48 (3):1718–1741. (Cited on page 13.)

Cattaneo, Matias D, Max H Farrell, Michael Jansson, and Ricardo Masini. 2024b. "Higher-order Refinements of Small Bandwidth Asymptotics for Density-Weighted Average Derivative Estimators." *arXiv preprint arXiv:2301.00277* . (Cited on page 32.)

Cattaneo, Matias D. and Michael Jansson. 2018. "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency." *Econometrica* 86 (3):955–995. (Cited on page 32.)

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2019. "Two-step Estimation and Inference with Possibly Many Included Covariates." *Review of Economic Studies* 86 (3):1095–1122. (Cited on page 32.)

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey. 2018. "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity." *Journal of the American Statistical Association* 113 (523):1350–1361. (Cited on pages 32 and 49.)

Chatla, Suneel Babu and Galit Shmueli. 2020. "A Tree-Based Semi-Varying Coefficient Model for the COM-Poisson Distribution." *Journal of Computational and Graphical Statistics* 29 (4):827–846. (Cited on page 13.)

Chen, Rong and Ruey Tsay. 1993. "Functional-coefficient autoregressive models." *Journal of the American Statistical Association* 88 (421):298–308. (Cited on page 44.)

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics*, *Handbook of Econometrics*, vol. 6B, edited by J.J. Heckman and E.E. Leamer, chap. 76. Elsevier. (Cited on pages 13, 14, and 28.)

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018a. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21 (1):C1–C68. (Cited on pages 3, 4, 14, 18, 30, 31, 32, 49, 53, 54, and 62.)

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018b. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." *arXiv preprint arXiv:1712.04802* . (Cited on page 10.)

Chernozhukov, Victor, Mert Demirer, Greg Lewis, and Vasilis Syrgkanis. 2019. "Semi-Parametric Efficient Policy Learning with Continuous Actions." In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett. 15065–15075. (Cited on page 49.)

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M. Robins. 2022a. "Locally Robust Semiparametric Estimation." *Econometrica* 90 (4):1501–1535. (Cited on pages 14, 15, 17, 32, and 54.)

Chernozhukov, Victor, Whitney Newey, Vıctor M Quintas-Martınez, and Vasilis Syrgkanis. 2022b. "Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests." In *International Conference on Machine Learning*. PMLR, 3901–3914. (Cited on page 17.)

Chernozhukov, Victor, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. 2024a. "Automatic Debiased Machine Learning via Riesz Regression." *arXiv:2104.14737* . (Cited on page 17.)

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh. 2022a. "Automatic debiased machine learning of causal and structural effects." *Econometrica* 90 (3):967–1027. (Cited on page 17.)

———. 2022b. "Debiased machine learning of global and local parameters using regularized Riesz representers." *The Econometrics Journal* 25 (3):576–601. (Cited on page 17.)

Chernozhukov, Victor, Whitney K. Newey, Rahul Singh, and Vasilis Syrgkanis. 2024b. "Adversarial Estimation of Riesz Representers." *arXiv preprint arXiv:2101.00009* . (Cited on page 17.)

Chetverikov, Denis, Andres Santos, and Azeem M. Shaikh. 2018. "The Econometrics of Shape Restrictions." *Annual Review of Economics* 10 (1):31–63. (Cited on page 33.)

Cleveland, William S, Eric Grosse, and William M Shyu. 1991. "Local regression models." In *Statistical models in S*, edited by J. M. Chambers and T. Hastie. Pacific Grove: Wadsworth and Brooks/Cole, 309–376. (Cited on page 44.)

Colangelo, Kyle and Ying-Ying Lee. 2020. "Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments." *arXiv:2004.03036* . (Cited on page 51.)

De Chaisemartin, Clément and Xavier d'Haultfoeuille. 2023. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey." *The Econometrics Journal* 26 (3):C1–C30. (Cited on page 10.)

Donald, Stephen Geoffrey. 1990. *Estimation of heteroskedastic limited dependent variable models*. Ph.D. thesis, University of British Columbia. URL https://open.library.ubc.ca/collections/ubctheses/831/items/1.0103893. (Cited on page 52.)

Fan, Jianqing and Wenyang Zhang. 2008. "Statistical methods with varying coefficient models." *Statistics and Its Interface* 1 (1):179–195. (Cited on page 13.)

Farrell, Max H. 2015. "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations." *arXiv:1309.4686, Journal of Econometrics* 189:1–23. (Cited on pages 4, 15, 30, 47, and 48.)

Farrell, Max H., Tengyuan Liang, and Sanjog Misra. 2021. "Deep Neural Networks for Estimation and Inference." *arXiv:1809.09953, Econometrica* 89 (1):181–213. (Cited on pages 3, 4, 28, 29, 42, 54, 55, 56, and 57.)

Foster, Dylan J and Vasilis Syrgkanis. 2023. "Orthogonal statistical learning." *The Annals of Statistics* 51 (3):879–908. (Cited on pages 4, 14, 18, and 30.)

Gallant, A Ronald and Douglas W Nychka. 1987. "Semi-nonparametric maximum likelihood estimation." *Econometrica* 55 (2):363–390. (Cited on pages 11 and 13.)

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. Cambridge: MIT Press. (Cited on page 12.)

Graham, Bryan S and Cristine Campos de Xavier Pinto. 2018. "Semiparametrically efficient estimation of the average linear regression function." *Journal of Econometrics,* forthcoming . (Cited on pages 46 and 49.)

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66 (2):315–331. (Cited on pages 46 and 47.)

———. 2004. "Functional restriction and efficiency in causal inference." *Review of Economics and Statistics* 84 (1):73–76. (Cited on page 47.)

Hanin, Boris. 2017. "Universal function approximation by deep neural nets with bounded width and relu activations." *arXiv preprint arXiv:1708.02691* . (Cited on page 28.)

Hastie, Trevor and Robert Tibshirani. 1993. "Varying-Coefficient Models." *Journal of the Royal Statistical Society, Series B* 55 (4):757–796. (Cited on page 44.)

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score." *Econometrica* 71 (4):1161–1189. (Cited on page 47.)

Hirshberg, David A. and Stefan Wager. 2019. "Augmented Minimax Linear Estimation." *arXiv:1712.00038* . (Cited on page 49.)

Huang, Jianhua Z and Haipeng Shen. 2004. "Functional coefficient regression models for non-linear time series: a polynomial spline approach." *Scandinavian Journal of Statistics* 31 (4):515–534. (Cited on page 63.)

Ichimura, Hidehiko and Whitney K Newey. 2022. "The influence function of semiparametric estimators." *Quantitative Economics* 13 (1):29–61. (Cited on pages 30 and 58.)

Imbens, Guido W., Whitney K. Newey, and Geert Ridder. 2007. "Mean-Squared-Error Calculations for Average Treatment Effects." *working paper* . (Cited on page 47.)

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An introduction to statistical learning*, vol. 112. Springer, 2 ed. (Cited on page 12.)

Javanmard, Adel and Andrea Montanari. 2014. "Confidence intervals and hypothesis testing for high-dimensional regression." *The Journal of Machine Learning Research* 15 (1):2869–2909. (Cited on page 18.)

Jorgensen, Murray A. 1993. "Influence functions for iteratively defined statistics." *Biometrika* 80:253–253. (Cited on page 47.)

Kennedy, Edward H. 2023. "Towards optimal doubly robust estimation of heterogeneous causal effects." *Electronic Journal of Statistics* 17 (2):3008–3049. (Cited on page 30.)

Kennedy, Edward H, Sivaraman Balakrishnan, James M Robins, and Larry Wasserman. 2024. "Minimax rates for heterogeneous causal effect estimation." *The Annals of Statistics* 52 (2):793–816. (Cited on page 30.)

Kennedy, Edward H, Zongming Ma, Matthew D McHugh, and Dylan S Small. 2017. "Nonparametric methods for doubly robust estimation of continuous treatment effects." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 79 (4):1229. (Cited on page 51.)

Keydana, Sigrid. 2023. *Deep Learning and Scientific Computing with R Torch*. Chapman and Hall/CRC. (Cited on page 12.)

Koh, Pang Wei and Percy Liang. 2017. "Understanding Black-Box Predictions via Influence Functions." In *International conference on machine learning*. PMLR, 1885–1894. (Cited on page 18.)

Li, Qi, Cliff J Huang, Dong Li, and Tsu-Tan Fu. 2002. "Semiparametric smooth coefficient models." *Journal of Business & Economic Statistics* 20 (3):412–422. (Cited on page 44.)

Ma, Shujie, Jeffrey S Racine, and Lijian Yang. 2015. "Spline regression in the presence of categorical predictors." *Journal of Applied Econometrics* 30 (5):705–717. (Cited on page 13.)

Ma, Xinwei and Jingshen Wang. 2020. "Robust inference using inverse probability weighting." *Journal of the American Statistical Association* 115 (532):1851–1860. (Cited on page 17.)

Mammen, Enno and Sara van de Geer. 1997. "Penalized quasi-likelihod estimation in partially linear models." *Annals of Statistics* 25 (3):1014–1035. (Cited on pages 16, 47, and 49.)

Maurer, Andreas. 2016. "A Vector-Contraction Inequality for Rademacher Complexities." In *Algorithmic Learning Theory*, edited by Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles. Cham: Springer International Publishing, 3–17. (Cited on page 55.)

McFadden, Daniel and Kenneth Train. 2000. "Mixed MNL models for discrete response." *Journal of applied Econometrics* 15 (5):447–470. (Cited on page 22.)

Nevo, Aviv. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica* 69 (2):307–342. (Cited on page 50.)

Newey, Whitney K. 1990. "Semiparametric efficiency bounds." *Journal of Applied Econometrics* 5 (2):99–135. (Cited on page 58.)

Newey, Whitney K. 1994. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica* 62 (6):1349–1382. (Cited on pages 18, 30, 47, 52, 57, and 58.)

Newey, Whitney K. and Daniel L. McFadden. 1994. "Large sample estimation and hypothesis testing." In *Handbook of Econometrics*, *Handbook of Econometrics*, vol. 4, edited by R. F. Engle and D. McFadden, chap. 36. Elsevier, 2111–2245. (Cited on pages 10, 16, 18, and 54.)

Newey, Whitney K and James M Robins. 2018. "Cross-fitting and fast remainder rates for semiparametric estimation." *arXiv preprint arXiv:1801.09138* . (Cited on page 14.)

Newey, Whitney K and Thomas M Stoker. 1993. "Efficiency of weighted average derivative estimators and index models." *Econometrica* 61 (5):1199–1223. (Cited on page 51.)

Nie, Xinkun and Stefan Wager. 2020. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *arXiv:1712.04912. Biometria,* forthcoming . (Cited on page 30.)

O'Hagan, Anthony. 1978. "Curve fitting and optimal design for prediction." *Journal of the Royal Statistical Society: Series B* 40 (1):1–24. (Cited on page 44.)

Okui, Ryo, Dylan S. Small, Zhiqiang Tan, and James M. Robins. 2012. "Doubly Robust Instrumental Variable Regression." *Statistica Sinica* 22 (1):173–205. (Cited on page 50.)

Olsen, Randall J. 1978. "Note on the uniqueness of the maximum likelihood estimator for the Tobit model." *Econometrica* 46 (5):1211–1215. (Cited on page 52.)

Padilla, Oscar Hernan Madrid, Wesley Tansey, and Yanzhen Chen. 2022. "Quantile regression with ReLU networks: Estimators and minimax rates." *Journal of Machine Learning Research* 23 (247):1–42. (Cited on page 28.)

Papke, Leslie E and Jeffrey M Wooldridge. 1996. "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates." *Journal of Applied Econometrics* 11 (6):619–632. (Cited on page 50.)

Powell, James L., James H. Stock, and Thomas M. Stoker. 1989. "Semiparametric Estimation of Index Coefficients." *Econometrica* 57 (6):1403–1430. (Cited on page 51.)

Racine, Jeff and Qi Li. 2004. "Nonparametric estimation of regression functions with both categorical and continuous data." *Journal of Econometrics* 119 (1):99–130. (Cited on page 13.)

Roberts, Daniel A, Sho Yaida, and Boris Hanin. 2022. *The principles of deep learning theory*. Cambridge, MA, USA: Cambridge University Press. (Cited on page 12.)

Robins, James, Lingling Li, Eric Tchetgen, and Aad van der Vaart. 2008. "Higher order influence functions and minimax estimation of nonlinear functionals." In *Probability and Statistics: Essays in Honor of David A. Freedman*, vol. 2, edited by Deborah Nolan and Terry Speed. Beachwood, Ohio, USA: Institute of Mathematical Statistics. (Cited on page 18.)

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427):846–866. (Cited on page 47.)

———. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90 (429):846–866. (Cited on page 47.)

Robinson, Peter M. 1988. "Root-n-consistent Semiparametric Regression." *Econometrica* 56 (4):931–954. (Cited on page 48.)

Schmidt-Hieber, Johannes. 2019. "Nonparametric regression using deep neural networks with ReLU activation function." *arXiv:1708.06633, Annals of Statistics*, forthcoming . (Cited on page 46.)

Shalit, Uri, Fredrik D. Johansson, and David Sontag. 2017. "Estimating individual treatment effect: generalization bounds and algorithms." In *Proceedings of the 34th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 70, edited by Doina Precup and Yee Whye Teh. PMLR, 3076–3085. (Cited on page 30.)

Spall, James C. 1986. "An approximation for analyzing a broad class of implicitly and explicitly defined estimators." *Communications in Statistics-Theory and Methods* 15 (12):3747–3762. (Cited on page 47.)

Stone, Charles J. 1982. "Optimal global rates of convergence for nonparametric regression." *The annals of statistics* :1040–1053. (Cited on page 30.)

Stone, Charles J., Mark H. Hansen, Charles Kooperberg, and Young K. Truong. 1997. "Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture." *The Annals of Statistics* 25 (4):1371–1470. (Cited on page 44.)

Tambwekar, Anuj, Anirudh Maiya, Soma Dhavala, and Snehanshu Saha. 2022. "Estimation and Applications of Quantiles in Deep Binary Classification." *IEEE Transactions on Artificial Intelligence* 3 (2):275–286. (Cited on page 28.)

van der Vaart, Aad. 1998. *Asymptotic Statistics*. Cambridge University Press. (Cited on page 58.)

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2):3–28. URL https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3. (Cited on page 4.)

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 30. (Cited on page 9.)

Velez, Amilcar. 2024. "On the Asymptotic Properties of Debiased Machine Learning Estimators." *arXiv preprint arXiv:2411.01864* . (Cited on page 32.)

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523):1228–1242. (Cited on page 10.)

Wooldridge, Jeffrey M. 2004. "Estimating average partial effects under conditional moment independence assumptions." *cemmap working paper CWP03/04* . (Cited on page 49.)

———. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 2 ed. (Cited on page 51.)

Yarotsky, Dmitry. 2017. "Error bounds for approximations with deep ReLU networks." *Neural Networks* 94:103–114. (Cited on pages 28 and 57.)

———. 2018. "Optimal approximation of continuous functions by very deep ReLU networks." *arXiv preprint arXiv:1802.03620* . (Cited on page 28.)

Zadrozny, Bianca. 2004. "Learning and evaluating classifiers under sample selection bias." ICML '04. New York, NY, USA: Association for Computing Machinery, 114. URL https://doi.org/10.1145/1015330.1015425. (Cited on page 5.)

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. 2008. "Model-based recursive partitioning." *Journal of Computational and Graphical Statistics* 17 (2):492–514. (Cited on page 13.)

Zhang, Cun-Hui and Stephanie S. Zhang. 2014. "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society. Series B* 76 (1):217–242. (Cited on page 18.)

Appendix A Implementation Details

A.1 Double Machine Learning

For simplicity, divide the observations $\{1, \dots, n\}$ into three subsets, which we denote $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$.[4] On sample $\mathcal{A}$ we solve (3.5) to obtain $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$. In principle, any structurally compatible nonparametric/ML method could be used here. We apply deep neural networks for the reasons already discussed, the most important for the present is the automatic differentiation engine.

Then in sample $\mathcal{B}$ we obtain $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{x})$. Here is the first use of automatic differentiation: from the optimization of (3.5), we can obtain the *values* $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}_i, \boldsymbol{t}_i, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, $i \in \mathcal{B}$, automatically. This is done automatically, without obtaining the $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\cdots)$ as a function per se. If the functional form of $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\cdots)$ is already known, then this could be applied to the data. But in many cases, this function is not known, and there is no need to characterize it since the values can be obtained automatically. In other words, $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}_i, \boldsymbol{t}_i, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, $i \in \mathcal{B}$ is simply a column of data and we project this nonparametrically on $\boldsymbol{x}_i$ to obtain an estimate of the conditional expectation $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}]$. Any regression method can be used here. We emphasize that this is *identical* to the more standard practice of obtaining the functional derivatives, characterizing the nonparametric object that is $\boldsymbol{\Lambda}(\boldsymbol{x})$, and then estimating it: we are not doing a numerical approximation.

Finally, in sample $\mathcal{C}$, we estimate $\boldsymbol{\mu}_0$. Again, we can use automatic differentiation to obtain the values of any unknown functions. For example, in our application in Section 5.1, the function $\boldsymbol{H}$ that defines $\boldsymbol{\mu}_0$ involves a fixed-point problem, and is not available in closed form. Thus the gradient piece of the influence function, $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \cdot; \boldsymbol{t}^*)$ is not available as a function. However, we can obtain

[4] In Appendix A we discuss general cross fitting and heuristics for splitting the sample.

the values of this derivative on the data, $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i); \boldsymbol{t}^*)$, $i \in \mathcal{C}$, automatically, which is all that is required. The same is true for $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$. Thus we can compute the sample average

$$\widehat{\boldsymbol{\mu}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i), \widehat{\boldsymbol{\Lambda}}(\boldsymbol{x}_i)), \tag{A.1}$$

where $|\mathcal{C}|$ is the number of observations in subset $\mathcal{C}$, and the functions $\widehat{\boldsymbol{\theta}}(\cdot)$ and $\widehat{\boldsymbol{\Lambda}}(\cdot)$ are learned on $\mathcal{A}$ and $\mathcal{B}$ as above. Theorem 2 verifies that $\widehat{\boldsymbol{\mu}}$ is asymptotically Normal with variance $\boldsymbol{\Psi} = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\Lambda})^2]$, which can be consistently estimated by

$$\widehat{\boldsymbol{\Psi}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \left(\boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{t}_i, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i), \widehat{\boldsymbol{\Lambda}}(\boldsymbol{x}_i)) - \widehat{\boldsymbol{\mu}} \right)^2. \tag{A.2}$$

For example, if $d_{\boldsymbol{\mu}} = 1$ so the parameter of interest is a scalar,

$$\left[\widehat{\mu} - 1.96\sqrt{\widehat{\Psi}/n} \ , \ \widehat{\mu} + 1.96\sqrt{\widehat{\Psi}/n} \right]$$

is a valid 95% confidence interval. Obtaining this interval in fully generic structural settings is no more difficult than current practice for standard cases, e.g. the average effect of a binary treatment.

In theory, the size of each subset is proportional to n and thus sample splitting does not impact convergence rates or precision, but in practice this can be a poor approximation. Farrell, Liang, and Misra (2021) show that for some estimands sample splitting is not needed for inference after deep learning (under essentially the same assumptions as here). It would be useful to extend that argument to more general settings.

A.2 SOMETHING ELSE?????

Tuning parameter choices

code

optimization

Appendix B Examples and Discussion

In this section we discuss a few special cases of our methodology to build intuition and connect to prior work, particularly other work on semiparametric inference. These examples help place our work in the literature, illustrate how familiar models can be enriched with deep learning, and show how the identification assumptions required on parametric models carry over.

B.1 Generalized Linear Models

One important special case, which covers many prior works in both nonparametric first steps and semiparametric contexts, is modeling the conditional mean of a scalar outcome. The workhorse model in this case would be

$$\mathbb{E}[Y \mid \boldsymbol{T} = \boldsymbol{t}] = G(\alpha_0 + \boldsymbol{\beta}_0' \boldsymbol{t}), \tag{B.1}$$

for an inverse link function $G(u) : \mathbb{R} \to \mathbb{R}$ (potentially along with other assumptions about variances or other parts of the data generating process). Standard examples include linear and logistic regression. Slope coefficients $\boldsymbol{\beta}_0$ in such models, along with marginal effects in nonlinear models, are among the most commonly studied objects in economic research.

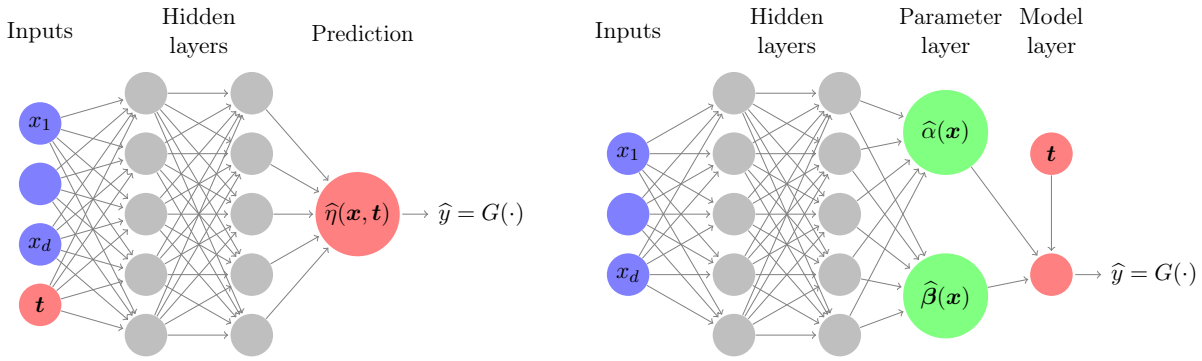
To enrich this model with ML, we change the intercept and slope to the parameter functions $\boldsymbol{\theta}(\boldsymbol{x}) = (\alpha_0(\boldsymbol{x}), \boldsymbol{\beta}_0(\boldsymbol{x}))'$ and assume that

$$\mathbb{E}[Y \mid \boldsymbol{T} = \boldsymbol{t}, \boldsymbol{X} = \boldsymbol{x}] = G(\alpha_0(\boldsymbol{x}) + \boldsymbol{\beta}_0(\boldsymbol{x})' \boldsymbol{t}). \tag{B.2}$$

This formulation maintains the structural relationship between $\boldsymbol{T}$ and Y but allows for rich heterogeneity in $\boldsymbol{X}$. In contrast to our structured approach, the naive ML approach would treat $\boldsymbol{X}$ and $\boldsymbol{T}$ as equally informative about Y, and assume, for an unknown $\eta(\boldsymbol{t}, \boldsymbol{x})$ to be estimated, that

$$\mathbb{E}[Y \mid \boldsymbol{T} = \boldsymbol{t}, \boldsymbol{X} = \boldsymbol{x}] = G(\eta(\boldsymbol{x}, \boldsymbol{t})). \tag{B.3}$$

Deep learning architectures to estimate these two models are shown in Figure B.1. Although both ultimately minimize the same prediction loss, the structural approach uses flexibility and expressive power of ML to learn the parameter functions in order to reduce the loss. The figure illustrates



(a) Standard prediction architecture

(b) Structured deep learning

Figure B.1: Comparing the standard prediction-focused architecture to learning parameter functions using the structured approach, matching the models of (B.3) and (B.2) respectively.

how easy it is to implement the structural approach, with only an extra line or two of code. As an aside, Figure B.1 also illustrates how our method (and theory) applies to generalized additive models, where the different components of $\boldsymbol{\theta}_0$ are known to rely on different subsets of $\boldsymbol{X}$: we simply sever the appropriate links, so that separate networks feed into the parameter layer nodes according to the model.

Interpreting $\boldsymbol{\beta}_0(\boldsymbol{x})$ in (B.2) is the same as interpreting the original (homogeneous) slope $\boldsymbol{\beta}_0$ in (B.1). In contrast, (B.3) is unstructured and uninterpretable. Here $\eta(\boldsymbol{x},\boldsymbol{t})$ is truly a nuisance function, where "nuisance" is taken literally to mean annoying and uninteresting. As demonstrated in Section 2, the lack of structure makes this output not useful. Further, $\eta(\boldsymbol{x},\boldsymbol{t})$ can only be learned at a much slower rate, governed by $d_{\boldsymbol{X}} + d_{\boldsymbol{T}}$, compared to learning $\boldsymbol{\beta}_0(\boldsymbol{x})$ which depends only upon $d_{\boldsymbol{X}}$. Though generally assumed away in theory since dimensions are finite, this difference matters in applications. All in all, it is better to view (B.2) as an enriched version of (B.1) rather than a restricted version of (B.3).

Remark B.1 (Origins of Model (B.2))**.** Models of the form (B.2) are common in the statistics literature, referred to as a "varying coefficient" model (Cleveland, Grosse, and Shyu, 1991; Hastie and Tibshirani, 1993), "functional coefficient" model (Chen and Tsay, 1993), or "smooth coefficient" model (Li et al., 2002), and falls into the class of "extended linear models" as in Stone et al. (1997). O'Hagan (1978) may be the earliest treatment. Our motivation is generally different from these literatures, but our results still apply to all these cases, as well as to generalized additive models. A further restriction, to constant slopes, yields partially linear models. See also Remarks B.2 and 6

and other discussion in Appendix B. ♣

We briefly summarize our theoretical results in the special case of an enriched generalized linear model. In this case, we can verify the high level conditions using the following familiar, primitive assumptions.

Assumption 4. *(i) The conditional expectation $G(\alpha_0(\boldsymbol{x}) + \boldsymbol{\beta}_0(\boldsymbol{x})'\boldsymbol{t})$ enters the loss through a known, real-valued transformation $g(\cdot)$, where (i) g and G are continuously invertible and $g/\|g\|_\infty$ and $G/\|G\|_\infty$ belong to $\mathcal{W}^{p,\infty}([-1,1])$, for $p \geq 3$. (ii) Assumption 1 holds with $\ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ replaced by $\ell(\boldsymbol{y}, g)$, and the conditions therein apply to the scalar argument g. (iii) The eigenvalues of $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X} = \boldsymbol{x}]$ are bounded and bounded away from zero uniformly in $\boldsymbol{x}$.*

Condition (i) ensures that the loss function is sufficiently smooth while (ii) and (iii) ensures the curvature through the standard positive variance condition. These conditions are familiar from the parametric case, where, for example, assuming that $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}']$ is positive definite would be standard.

Specializing Theorem 1 to this case, we have the following result.

Corollary B.1. *Let the conditions of Theorem 1 and Assumption 4 hold. Then for a DNN structured according to Figure B.1, $\|\widehat{\theta}_k - \theta_{0k}\|^2_{L_2(\boldsymbol{X})} = O(n^{-\frac{p}{p+d_c}} \log^8(n))$, for $k = 1, \ldots, d_{\boldsymbol{\theta}}$, and $\|G(\widehat{\alpha}(\boldsymbol{x}) + \widehat{\boldsymbol{\beta}}(\boldsymbol{x})'\boldsymbol{t}) - G(\alpha_0(\boldsymbol{x}) + \boldsymbol{\beta}_0(\boldsymbol{x})'\boldsymbol{t})\|^2_{L_2(\boldsymbol{X})} = O(n^{-\frac{p}{p+d_c}} \log^8(n))$.*

Here we give two results. First, we show that we can estimate the heterogeneous intercept and slope parameters at the appropriate rate, depending on the dimension of the heterogeneity. This is direct from Theorem 1. This is required as economic constructs depend on these parameters, rather than on the conditional expectation function as a whole. But we also give a result for the mean function in the enriched model (B.2). This result may be of independent interest, as it establishes that structural deep learning has good statistical properties in varying coefficient models, additive models, and other such cases. It is also useful for comparing to the more typical use of inference after ML, where the (prediction) function $\mathbb{E}[Y|\boldsymbol{x}, \boldsymbol{t}]$ would be unstructured and would be learnable at a rate dependent on $d_{\boldsymbol{X}} + d_{\boldsymbol{T}}$. For example, it is much easier (statistically) to estimate $\mathbb{E}[\boldsymbol{\beta}_0(\boldsymbol{X})]$ than the average derivatives $\mathbb{E}[\partial G(\eta(\boldsymbol{X}, \boldsymbol{T}))/\partial \boldsymbol{t}]$, even though both are linear summaries of the dependence of Y on $\boldsymbol{T}$.

Remark B.2 (Adaptivity)**.** Specific forms of neural networks are known to be adaptive to structures like (B.2) in the sense that if the structure holds, then even if estimation is done under the generic (B.3), the estimator will still obtain the same rate as though (B.2) was imposed (Bach, 2017; Bauer and Kohler, 2019; Schmidt-Hieber, 2019). This is perhaps not surprising, since neural networks are based on compositions and (B.2) is based on a composition of simpler functions. This is still not useful for our purpose because $\boldsymbol{\theta}(\boldsymbol{x}) = (\alpha_0(\boldsymbol{x}), \boldsymbol{\beta}_0(\boldsymbol{x}))'$ cannot be recovered from an adaptive procedure and used in the second stage. The estimator *statistically* adapts to the structure, but not *economically*. Economic structure is enforced on the data, not discovered from the data. Further, experience shows that, despite the (asymptotic) theory of adaptivity, finite sample performance is improved by structure. ♣

B.1.1 Inference

Specializing our influence function to case of (B.2) helps connect with past work. To save notation, let $\widetilde{\boldsymbol{T}} = (1, \boldsymbol{T}')'$ and define $G_0(\boldsymbol{x}, \boldsymbol{t}) = G(\alpha_0(\boldsymbol{x}) + \boldsymbol{\beta}_0(\boldsymbol{x})'\boldsymbol{t})$ and $\dot{G}_0(\boldsymbol{x}, \boldsymbol{t}) = dG(u)/du$ at $u = \alpha_0(\boldsymbol{x}) + \boldsymbol{\beta}_0(\boldsymbol{x})'\boldsymbol{t}$. If we take the standard approach where $\boldsymbol{\ell}_{\boldsymbol{\theta}}(y, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) = \widetilde{\boldsymbol{t}}(G_0(\boldsymbol{x}, \boldsymbol{t}) - y)$, the influence function is

$$\boldsymbol{\psi}(y, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*\right) + \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)\mathbb{E}[\dot{G}_0(\boldsymbol{x}, \boldsymbol{T})\widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{T}}' \mid \boldsymbol{X} = \boldsymbol{x}]^{-1}\widetilde{\boldsymbol{t}}(y - G_0(\boldsymbol{x}, \boldsymbol{t})). \quad \text{(B.4)}$$

Further, if $d_{\boldsymbol{T}} = d_{\boldsymbol{\mu}} = 1$, the above result simplifies to

$$\begin{aligned}\psi(y, t, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = {} & H\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); t^*\right) \\ & + \frac{H_\alpha(\boldsymbol{x})\left(\lambda_2(\boldsymbol{x}) - \lambda_1(\boldsymbol{x})t\right) + H_\beta(\boldsymbol{x})\left(\lambda_0(\boldsymbol{x})t - \lambda_1(\boldsymbol{x})\right)}{\lambda_2(\boldsymbol{x})\lambda_0(\boldsymbol{x}) - \lambda_1(\boldsymbol{x})^2}\Big(y - G_0(\boldsymbol{x}, \boldsymbol{t})\Big),\end{aligned} \quad \text{(B.5)}$$

where $\lambda_k(\boldsymbol{x}) = \mathbb{E}[\dot{G}_0(\boldsymbol{x}, \boldsymbol{T})T^k | \boldsymbol{X} = \boldsymbol{x}], k = \{0, 1, 2\}$, $H_\alpha(\boldsymbol{x}) = \partial H(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)/\partial\alpha$, and $H_\beta(\boldsymbol{x}) = \partial H(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)/\partial\beta$. In nonlinear models the need for three-way sample splitting is clear: the regressand of $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\dot{G}_0(\boldsymbol{x}, \boldsymbol{T})\widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{T}}' \mid \boldsymbol{X} = \boldsymbol{x}]$ depends on unknown $\boldsymbol{\theta}_0(\boldsymbol{x})$, which must be estimated in a first step. For linear models, three splits are not necessary.

This result recovers several known influence functions. For linear models, it matches Hahn (1998) for binary treatments and Graham and Pinto (2018) for in the continuous case. For mean square

projections in general, see Newey (1994). For semiparametric regression, with constant $\boldsymbol{\beta}_0$, our result gives valid inference but does not match the efficient influence function (Mammen and van de Geer, 1997) because we do not impose constancy.

We can also effortlessly obtain new results. For example, in Section 5.1 we conduct inference on the fully flexible average marginal effects in a logistic regression, where $\boldsymbol{\mu}_0 = \mathbb{E}[\dot{G}(\alpha_0(\boldsymbol{X}) + \boldsymbol{\beta}_0(\boldsymbol{X})'\boldsymbol{T})\boldsymbol{\beta}_0(\boldsymbol{X})]$. In this case the function $\boldsymbol{H}$ and its derivatives are available in closed form, and the derivatives of the loss are well known, but again, this is not necessary. We also use (B.4) for a function $\boldsymbol{H}$ that is not available in closed form (Spall, 1986; Jorgensen, 1993).

B.2 Average Effect of a Binary Treatment

Perhaps the most well known and commonly used influence function is for the average effect of a binary treatment in observational data. The history of this influence function is illustrative: it was characterized precisely first for the purposes of efficiency considerations (Hahn, 1998, 2004), later used to show that certain plug-in estimators could be efficient (Hirano, Imbens, and Ridder, 2003; Imbens, Newey, and Ridder, 2007) under strong assumptions, then recently used to obtain valid inference under weak first-stage conditions (Cattaneo, 2010; Farrell, 2015). This influence function matches the doubly robust estimator (Robins, Rotnitzky, and Zhao, 1994, 1995) and has been used in other contexts to obtain sharper results, see Remarks 6 and 4.

Here we have a scalar outcome and $\boldsymbol{T} = T = \{0, 1\}$ is the scalar binary treatment indicator. Let $Y(t)$ be the potential outcome under treatment $T = t$ and assume that the observed outcome $Y = TY(1) + (1-T)Y(0)$ along with unconfoundedness. The parameter of interest is $\mu_0 = \mathbb{E}[Y(1) - Y(0)]$. In a randomized experiment without additional covariates, one estimates μ_0 with a difference in means, which is equivalent to running a regression of the observed outcome Y on the dummy variable T. The enriched structural version is then (B.2), with $G(u) = u$, so that $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \alpha_0(\boldsymbol{x}) + \beta_0(\boldsymbol{x})\boldsymbol{t}$. In this notation, $\beta_0(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{x}]$ is the conditional average treatment effect (CATE) function.

The average effect $\mu_0 = \mathbb{E}[\beta_0(\boldsymbol{X})]$ is defined via $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{\theta}_0; \boldsymbol{t}^*) = \beta_0$. Equation (B.5) matches Hahn (1998), because in this case $H_\alpha = 0$, $H_\beta = 1$, $\lambda_0 = 1$, and $\lambda_1(\boldsymbol{x}) = \lambda_2(\boldsymbol{x}) = \mathbb{P}[T = 1 | \boldsymbol{X} = \boldsymbol{x}] := p(\boldsymbol{x})$, the propensity score, and so by adding and subtracting $p(\boldsymbol{x})$ and using the fact that $(1-t)t = 0$, we

have

$$\begin{aligned}
\psi(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\Lambda}) &= \theta_{02}(\boldsymbol{x}) + \frac{\dot{H}_1(\boldsymbol{x})(\lambda_2(\boldsymbol{x}) - \lambda_1(\boldsymbol{x})t) + \dot{H}_2(\boldsymbol{x})(\lambda_0(\boldsymbol{x})t - \lambda_1(\boldsymbol{x}))}{\lambda_2(\boldsymbol{x})\lambda_0(\boldsymbol{x}) - \lambda_1(\boldsymbol{x})^2}(y - G(\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t})) \\
&= \theta_{02}(\boldsymbol{x}) + \frac{(t - p(\boldsymbol{x}))(y - \theta_{01}(\boldsymbol{x}) - \theta_{02}(\boldsymbol{x})t)}{p(\boldsymbol{x}) - p(\boldsymbol{x})^2} \\
&= \theta_{02}(\boldsymbol{x}) + \frac{[(1 - p(\boldsymbol{x}))t - p(\boldsymbol{x})(1 - t)](y - \theta_{01}(\boldsymbol{x}) - \theta_{02}(\boldsymbol{x})t))}{p(\boldsymbol{x})(1 - p(\boldsymbol{x}))} \\
&= \theta_{02}(\boldsymbol{x}) + \frac{(1 - p(\boldsymbol{x}))t(y - \theta_{01}(\boldsymbol{x}) - \theta_{02}(\boldsymbol{x})t))}{p(\boldsymbol{x})(1 - p(\boldsymbol{x}))} - \frac{p(\boldsymbol{x})(1 - t)(y - \theta_{01}(\boldsymbol{x}) - \theta_{02}(\boldsymbol{x})t))}{p(\boldsymbol{x})(1 - p(\boldsymbol{x}))} \\
&= \theta_{02}(\boldsymbol{x}) + \frac{t(y - \theta_{01}(\boldsymbol{x}) - \theta_{02}(\boldsymbol{x})t))}{p(\boldsymbol{x})} - \frac{(1 - t)(y - \theta_{01}(\boldsymbol{x})))}{(1 - p(\boldsymbol{x}))}.
\end{aligned}$$

In this example, the standard overlap assumption, that the propensity score is bounded away from zero and one, ensures that $\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}$ is well behaved: the determinant of $\boldsymbol{\Lambda}(\boldsymbol{x}) = p(\boldsymbol{x})(1 - p(\boldsymbol{x}))$, the initial denominator above.

It is straightforward to extend this example in a number of directions. Additional mean parameters could be added to cover average treatment effects for specific treatment groups or multi-valued treatments (see Cattaneo (2010) and Cattaneo and Farrell (2011) for inference using classical nonparametrics (series) and Farrell (2015) for machine learning (group lasso) results). Beyond means, we could consider the variance of $Y(1)$ versus that of $Y(0)$ by taking a quasi-likelihood approach, letting $\ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ include the variance instead of simply fitting least squares. The conditions for convexity of the loss are well-known from likelihood theory and can be directly used here.

B.3 Partially Linear Models

Partially linear models are a common case for semiparametric inference, dating to the seminal work of Robinson (1988). Here (B.2) holds, but with $\boldsymbol{\beta}_0(\boldsymbol{x}) = \boldsymbol{\beta}_0$ constant. Restricting the slope to be constant rules out all treatment effect heterogeneity and is a strong assumption that should be used with caution.

For simplicity, consider the case with a scalar treatment variable. If β_0 is the parameter of interest, (B.5) shows that $\psi(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{\Lambda}) - \beta_0$ is

$$\left[\lambda_2(\boldsymbol{x}) - \frac{\lambda_1(\boldsymbol{x})^2}{\lambda_0(\boldsymbol{x})}\right]^{-1}\left(t - \frac{\lambda_1(\boldsymbol{x})}{\lambda_0(\boldsymbol{x})}\right)\Big(y - G(\alpha_0(\boldsymbol{x}) + \beta_0 t)\Big).$$

We must assume that $\lambda_2(\boldsymbol{x})\lambda_0(\boldsymbol{x}) \neq \lambda_1(\boldsymbol{x})^2$, which for identity G requires positive conditional variance of T. In nonlinear models the conditional moments will be weighted by $\dot{G}$ if we have used the appropriate loss. In some cases the nonsingularity will follow from other regularity conditions, such as for the logistic link, where $\dot{G} = G(1-G)$ and the Hessian is invertible under bounded covariates and we use the log-likelihood.

Partially linear models have featured prominently in the recent literature on high dimensional and ML settings, particularly with idenity link. The pioneering work of Belloni, Chernozhukov, and Hansen (2014) established valid inference after lasso selection in this context. Chernozhukov et al. (2018a) use it as the leading example of their generic results, and present several different Neyman orthogonal scores that could be used, none of which agree with ours due to the fact that we do not impose constant slope. Cattaneo, Jansson, and Newey (2018) study high dimensional linear models, including many-terms series settings, establish a more refined distributional approximation, and study standard error (in)validity. For the case of nonlinear link function, Carroll et al. (1997) and Mammen and van de Geer (1997) are closest to our work, while Belloni, Chernozhukov, and Wei (2016) study high-dimensional sparse models. Not imposing the constant slope means we do not attain the efficiency bound in general, cf Mammen and van de Geer (1997), but only under restrictions on the variance, such as Y being homoskedastic in $\boldsymbol{T}$ in the linear case.

Finally, our results could be used to study other components of the partially linear setting. For example, in both empirical finance Cattaneo et al. (2020) and applied microeconomics Cattaneo et al. (2024a) the function $\alpha(\boldsymbol{x})$ is of direct interest and we could conduct inference on the average.

B.4 Continuous Treatments, Average Effects, Fractional Outcomes

Comparatively less work has been done for the case of the generalized linear model (B.2). Wooldridge (2004) and Graham and Pinto (2018) consider the linear model case and discuss conditions for a causal interpretation of $\mathbb{E}[\boldsymbol{\beta}_0(\boldsymbol{X})]$. Our result in (B.4) matches the locally efficient influence function of Graham and Pinto (2018). Hirshberg and Wager (2019) use a different approach to recover the average effect, but briefly discuss double robustness. Chernozhukov et al. (2019) use the model with the goal of policy targeting.

Our method could be used to enrich other structural models in this context. As a first example, consider the so-called Berry logit (Berry, 1994) model for demand. Here the outcome is the market

share distribution across firms. The researcher observes $\{Y_{jm}\}$ which represent a collection of $j = 0 \dots J$ market shares across $m = 1 \dots M$ markets. The objective is then to model these as a function of firm (marketing) decisions $\boldsymbol{T}_{jm}$ (see e.g. Nevo (2001)). We can introduce heterogeneity across markets by allowing for the marketing effects to be moderated by consumer characteristics $\boldsymbol{x}_m$, so that we can write a collection of $(J-1)$ equations as follows

$$\mathbb{E}\left[\log\left(\frac{Y_{jm}}{Y_{0m}}\right)\middle|\, \boldsymbol{X} = \boldsymbol{x}_m, \boldsymbol{T}_{jm} = \boldsymbol{t}_{jm}\right] = \alpha_{0j}(\boldsymbol{x}_m) + \boldsymbol{\beta}_0(\boldsymbol{x}_m)'(\boldsymbol{t}_{jm} - \boldsymbol{t}_{0m}).$$

Stacking these equations and the corresponding data allows us to construct an estimator for $\alpha_{0j}(\boldsymbol{x}_m)$ and $\boldsymbol{\beta}_0(\boldsymbol{x}_m)$. We could extend this to include instruments (Okui et al., 2012).

Second, consider the Cobb-Douglas production function with heterogeneous parameters, which is given by $Y = CK^{\theta_{01}(\boldsymbol{x})}L^{\theta_{02}(\boldsymbol{x})}$. The standard approach is to take logs and fit

$$\mathbb{E}\left[\log Y \mid \boldsymbol{X} = \boldsymbol{x}, K = k, L = l\right] = \log C + \theta_{01}(\boldsymbol{x}) \log k + \theta_{02}(\boldsymbol{x}) \log l.$$

We can then estimate the structural parameters using our deep learning approach and conduct inference to decide if on average the technology exhibits increasing, constant, or decreasing returns to scale by computing $\boldsymbol{\mu}_0 = \mathbb{E}[\theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x})]$. The Cobb-Douglas specification has also been used in demand settings and marketing mix models and the framework described above would be readily applicable there as well.

Beyond linear cases, consider the fractional outcome model where the outcome Y is continuous but restricted to lie in $[0, 1]$. Following the seminal treatment of Papke and Wooldridge (1996), we take a quasi-likelihood approach, using logistic distribution with mean given by Equation (B.2). Papke and Wooldridge (1996) explicitly advocate the use of structure to ensure that the outcomes remain on the unit interval and argue that this specification is valid even at the endpoints and is more practically relevant then transformations of the dependent variable.

In the application of Papke and Wooldridge (1996), the data is at the firm level, the outcome is 401(k) participation, and the policy variable is the firm's rate of contribution matching. The quantities of interest are the marginal effect of the match rate on participation and the degree to which this marginal effect exhibits diminishing patterns. Our framework makes it trivial to enrich

this model with heterogeneity across firms and then conduct inference on average marginal effects or the average change in the marginal effect, given by

$$\text{AME}\left(t^*\right) = \mathbb{E}\left[\left.\frac{\partial \mathbb{E}[Y \mid \mathbf{X}, t]}{\partial t}\right|_{t=t^*}\right] \quad \text{and} \quad \text{ACME}\left(t^*\right) = \mathbb{E}\left[\left.\frac{\partial^2 \mathbb{E}[Y \mid \mathbf{X}, t]}{\partial t^2}\right|_{t=t^*}\right].$$

Because of the structure of the model, these are easily recovered in the form of $\boldsymbol{\mu}_0$, by taking $H_{\text{AME}}(\boldsymbol{x}, \boldsymbol{\theta}; t^*) = \beta_0 G_0(\boldsymbol{x}, \boldsymbol{t})(1 - G_0(\boldsymbol{x}, \boldsymbol{t}))$ and $H_{\text{ACME}}(\boldsymbol{x}, \boldsymbol{\theta}; t^*) = \beta_0^2 G_0(\boldsymbol{x}, \boldsymbol{t})(1 - G_0(\boldsymbol{x}, \boldsymbol{t}))(1 - 2G_0(\boldsymbol{x}, \boldsymbol{t}))$, respectively.

It is useful to contrast our model with the unstructure, naive ML approach. We impose structure on the nature of the effect of $\boldsymbol{T}$ on $\boldsymbol{Y}$. For unrestricted effects of continuous treatments using influence functions, see Kennedy et al. (2017) and Colangelo and Lee (2020). The unrestricted model may increase the generality of the results but can make inference and interpretation more difficult. A common parameter of interest in such cases is the average derivative (Powell, Stock, and Stoker, 1989; Newey and Stoker, 1993). This represents the average of a linear approximation of an unstructured relationship of $\boldsymbol{T}$ to Y. Our approach is perhaps more direct and transparent: if a linear approximation is of interest in the end, we directly enrich the linear approximation, rather than recover it from a more complex object. This contrast can also be seen in the fractional outcome models, where recovering the second derivative of a complex $G(\widehat{\eta}(\boldsymbol{x}, \boldsymbol{t}))$ could be challenging but our approach is transparent and simple.

B.5 Tobit

The type I Tobit model is well-studied in the parametric case, and so it serves as a useful example to illustrate how existing knowledge can be used to interpret the assumptions required in the enriched case.

The observed outcome is $Y = \max(0, Y^*)$, where in the parametric, homogeneous case Y^* is Gaussian given $\boldsymbol{T} = \boldsymbol{t}$ with mean $\alpha_0 + \boldsymbol{\beta}_0' \boldsymbol{t}$ and variance σ^2. The enriched version has mean $\alpha_0(\boldsymbol{x}) + \boldsymbol{\beta}_0(\boldsymbol{x})' \boldsymbol{t}$ and variance $\sigma^2(\boldsymbol{x})$ (in practice it can help to fit $\sigma^2(\boldsymbol{x}) = \exp\{\tilde{\sigma}(\boldsymbol{x})\}$). As usual, we work with the transformed parameters $\boldsymbol{\theta}_0(\boldsymbol{x}) = (\boldsymbol{\theta}_{01}(\boldsymbol{x})', \theta_{02}(\boldsymbol{x}))'$, with $\boldsymbol{\theta}_{01}(\boldsymbol{x}) = (\alpha_0(\boldsymbol{x}), \boldsymbol{\beta}_0(\boldsymbol{x})')'/\sigma(\boldsymbol{x})$ and $\theta_{02}(\boldsymbol{x}) = \sigma^{-1}(\boldsymbol{x})$. See Amemiya (1985) and Wooldridge (2010) for textbook treatments.

The gradient and Hessian are cumbersome but known. These can be used both for understanding

the assumptions required but also, if desired, in the computation. Let $\mathbb{1}_0 = \mathbb{1}\{Y^* \leq 0\}$ and $\mathbb{1}_1 = \mathbb{1}\{Y^* > 0\}$. Let ϕ and Φ denote the Gaussian density and distribution functions. With $\widetilde{\boldsymbol{t}} = (1, \boldsymbol{t}')'$, the gradient (score) terms are

$$\boldsymbol{\ell}_{\boldsymbol{\theta}_1}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = \mathbb{1}_0 \frac{\phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})\widetilde{\boldsymbol{t}}}{1 - \Phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})} - \mathbb{1}_1\Big(\theta_2(\boldsymbol{x})y - \boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}}\Big)\widetilde{\boldsymbol{t}}'$$

and

$$\boldsymbol{\ell}_{\theta_2}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = -\mathbb{1}_1\theta_2(\boldsymbol{x})^{-1} + \mathbb{1}_1(\theta_2(\boldsymbol{x})y - \boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})y.$$

The second derivatives are

$$\boldsymbol{\ell}_{\boldsymbol{\theta}_1\boldsymbol{\theta}_1}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = -\mathbb{1}_0 \frac{\phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})\widetilde{\boldsymbol{t}}\widetilde{\boldsymbol{t}}'}{1 - \Phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})} + \mathbb{1}_0 \frac{\phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})^2\widetilde{\boldsymbol{t}}\widetilde{\boldsymbol{t}}\widetilde{\boldsymbol{t}}'}{[1 - \Phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})]^2} + \mathbb{1}_1\widetilde{\boldsymbol{t}}\widetilde{\boldsymbol{t}}',$$

$$\ell_{\theta_2\theta_2}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = -\mathbb{1}_1\theta_2(\boldsymbol{x})^{-2} + \mathbb{1}_1 y^2, \qquad \text{and} \qquad \boldsymbol{\ell}_{\theta_1\theta_2}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = \mathbb{1}_1 y\widetilde{\boldsymbol{t}}.$$

That the gradients are conditionally mean zero can be directly verified. The matrix $\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}$ exists because $\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}} - \phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})/[1 - \Phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}})] > 0$, using exactly the logic from parametric models (Donald, 1990; Olsen, 1978; Amemiya, 1985). The remaining assumptions required would be standard for nonparametrics/ML, such as smoothness and boundedness.

B.6 Instrumental Variables

For multiple first stage objects, the influence function correction terms are generally additive (Newey, 1994) and our framework of course has the same feature. A good example is when $\boldsymbol{T}$ is endogenous and instrumental variables are available. Suppose there is single endogenous variable T and a single instrument Z. The standard approach would be two stage least squares. We enrich this to allow for fully flexible observed heterogeneity in the effects of the instrument and the endogenous variable, arriving at the two-equation model

$$Y = \theta_{01}(\boldsymbol{X}) + \theta_{02}(\boldsymbol{X})T + V, \tag{B.6}$$

$$T = \zeta_{01}(\boldsymbol{X}) + \zeta_{02}(\boldsymbol{X})Z + U, \tag{B.7}$$

where $\mathbb{E}[V \mid \boldsymbol{X}, Z] = \mathbb{E}[U \mid \boldsymbol{X}, Z] = 0$. For estimation, and moreover, derivation of an orthogonal score, we simply plug (B.7) into (B.6) to obtain the reduced form equation

$$\begin{aligned} &Y = \alpha_0(\boldsymbol{X}) + \beta_0(\boldsymbol{X})Z + \tilde{V}, \\ &\alpha_0(\boldsymbol{x}) = \theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x})\zeta_{01}(\boldsymbol{x}), \quad \beta_0(\boldsymbol{x}) = \theta_{02}(\boldsymbol{x})\zeta_{02}(\boldsymbol{x}), \quad \tilde{V} = \theta_{02}(\boldsymbol{X})U + V. \end{aligned} \tag{B.8}$$

This approach directly generalizes two-stage least squares to handle high-dimensional, complex heterogeneity, but notice that the linearity structure is maintained. To estimate the first stage parameter functions we simply apply (3.3) where the loss is the sum of two squared losses and the architecture in Figure B.1 is used twice. The leading case for inference would be the average partial effect $\mu_0 = \mathbb{E}[\theta_{02}(\boldsymbol{X})] = \mathbb{E}[\beta_0(\boldsymbol{X})/\zeta_{02}(\boldsymbol{X})]$. We again see the familiarity of the assumptions required: we need strong instruments, which here means we need $\zeta_{02}(\boldsymbol{X})$ to be nowhere zero. Any other function $H(\boldsymbol{X}, \alpha_0, \beta_0, \zeta_{01}, \zeta_{02}; \boldsymbol{t}^*)$ could be used.

The influence function will be of the standard form in this case, with essentially two copies of the linear case above. Define $\boldsymbol{\theta} = (\alpha_0, \beta_0, \zeta_{01}, \zeta_{02})'$, $\boldsymbol{w} = (y, t, z)$, $\widetilde{t} = (1, t)'$, $\widetilde{z} = (1, z)'$ and $\boldsymbol{I}_2$ the 2×2 identity matrix. Then

$$\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = -\begin{pmatrix} y - \alpha_0(\boldsymbol{x}) - \beta_0(\boldsymbol{x})z \\ t - \zeta_{01}(\boldsymbol{x}) - \zeta_{02}(\boldsymbol{x})z \end{pmatrix} \otimes \boldsymbol{z} \quad \text{and} \quad \boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = \boldsymbol{I}_2 \otimes \widetilde{z}\widetilde{z}'.$$

Therefore $\boldsymbol{\Lambda}(\boldsymbol{x}) = \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_Z(\boldsymbol{x})$, where $\boldsymbol{\Lambda}_Z(\boldsymbol{x}) = \mathbb{E}[\widetilde{z}\widetilde{z}' \mid \boldsymbol{X} = \boldsymbol{x}]$.

As before, our score is not the only possibility for estimation and inference IV models. Our approach aims for ease of use and transparency, both by sticking to the two stage least squares and by the structural compatibility of deep learning. Restricting to homogeneous effects, Chernozhukov et al. (2018a) study partially linear IV models and study three different possibly orthogonal scores, each requiring different functions in the first step and in the correction term. The same could be done in our unrestricted model. It is not obvious which score is best.

B.7 Two Step GMM

In Section 3 we focused on the case where the final parameter of interest was both just identified and available in closed form, that is, $\boldsymbol{\mu}_0 = \mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}), \boldsymbol{t}^*)]$ as in Equation (3.4). However, our first-step correction can also be used in GMM settings. In this case, the parameter of interest would be defined by the set of moment conditions $\mathbb{E}[\tilde{\boldsymbol{H}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}), \boldsymbol{\mu}_0, \boldsymbol{t}^*)] = 0$ for some $\tilde{\boldsymbol{H}}$. If the first step

was parametric, as in (3.1), then this would be standard two-step GMM and inference would follow from Newey and McFadden (1994, Theorem 6.1). In the enriched case the first step correction factor shares the same format, but is nonparametric in nature (just as before), and is given by (cf (4.1))

$$\phi(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\Lambda}, \boldsymbol{\theta}) = \tilde{\boldsymbol{H}}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})).$$

Following Chernozhukov et al. (2022a), adding this correction to the original moments yields the orthogonal moment conditions $\mathbb{E}[\tilde{\boldsymbol{H}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}), \boldsymbol{\mu}_0, \boldsymbol{t}^*) - \phi(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\Lambda}, \boldsymbol{\theta})] = 0$. Chernozhukov et al. (2022a), extending Chernozhukov et al. (2018a), show that the advantages of DML carry over to GMM based on these moments. Inference in our case follows directly from their results. Forming standard errors will require estimation of $\partial\mathbb{E}[\tilde{\boldsymbol{H}}(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{\mu}_0, \boldsymbol{t}^*)]/\partial\boldsymbol{\mu}$. See Chernozhukov et al. (2022a) for details, including discussion of cross-fitting, efficient weighting, and preliminary estimation. As before, our contribution here is methodological, not theoretical per se. Our first step correction factor applies quite generically and again, automatic differentiation can be used if necessary.

Appendix C Proofs

C.1 Bounds Structural Deep Learning – Theorem 1

Here we prove Theorem 1, using the proof method of Farrell, Liang, and Misra (2021). Some details will be deferred to that paper.

Define $\boldsymbol{\theta}_n \in \mathcal{F}_{\text{DNN}}$ as the best approximation to $\boldsymbol{\theta}_0$ in the class of DNNs as defined in the theorem statement. Let ϵ_n denote the error of the approximation:

$$\boldsymbol{\theta}_n = \underset{\boldsymbol{\theta} \in \mathcal{F}_{\text{DNN}}}{\arg\min} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty, \qquad \epsilon_n = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|_\infty.$$

Under Assumption 2 this error is controlled by the width and depth of an MLP, and we specify to this case at the end of the proof. For now we leave the error generic to allow for other approximation assumptions (such as other smoothness classes) and other architectures.

By Assumption 1 and that $\widehat{\boldsymbol{\theta}}$ optimizes ℓ over $\mathcal{F}_{\text{DNN}}$ in the data,

$$c_1\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))] - \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))] \\
&\leq \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))] - \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))] - \mathbb{E}_n[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))] + \mathbb{E}_n[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_n(\boldsymbol{X}))] \\
&= (\mathbb{E}-\mathbb{E}_n)\left[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X})) - \ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))\right] + \mathbb{E}_n\left[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_n(\boldsymbol{X})) - \ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))\right].
\end{aligned}
$$

Applying Farrell, Liang, and Misra (2021, Equation (A.2)) to the second term of the last line above, we find that with probability $1-e^{-\gamma}$

$$
\begin{aligned}
&c_1\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] \\
&\qquad \leq (\mathbb{E}-\mathbb{E}_n)\left[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X})) - \ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))\right] + c_2\epsilon_n^2 + \epsilon_n\sqrt{\frac{2C_\ell^2\gamma}{n}} + \frac{7C_\ell M\gamma}{n}. \quad \text{(C.1)}
\end{aligned}
$$

We now apply the localization-based analysis of Farrell, Liang, and Misra (2021) to the first term above and then collect the results. Suppose that for some r_0, $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2]^{1/2} \leq r_0$, which can always be attained given the boundedness. Let $\mathcal{F}^0_{\text{DNN}}$ be the subset of $\mathcal{F}_{\text{DNN}}$ such that $\boldsymbol{\theta} \in \mathcal{F}^0_{\text{DNN}}$ if $\mathbb{E}[\|\boldsymbol{\theta}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2]^{1/2} \leq r_0$. Then by Theorem 2.1 in Bartlett, Bousquet, and Mendelson (2005), for $\mathcal{G} = \{g = \ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}(\boldsymbol{x})) - \ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}_0(\boldsymbol{x})) : \boldsymbol{\theta} \in \mathcal{F}^0_{\text{DNN}}\}$, we find that, with probability at least $1-2e^{-\gamma}$, the empirical process term of (C.1) is bounded as

$$
(\mathbb{E}-\mathbb{E}_n)\left[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X})) - \ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))\right] \leq 6\mathbb{E}_\eta R_n\mathcal{G} + \sqrt{\frac{2C_\ell^2 r_0^2\gamma}{n}} + \frac{23\cdot 3MC_\ell}{3}\frac{\gamma}{n}, \quad \text{(C.2)}
$$

where

$$
R_n\mathcal{G} = \sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \eta_i g(\boldsymbol{w}_i) = \sup_{\boldsymbol{\theta}\in\mathcal{F}^0_{\text{DNN}}}\frac{1}{n}\sum_{i=1}^n \eta_i\left(\ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}(\boldsymbol{x})) - \ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}_0(\boldsymbol{x}))\right).
$$

is the empirical Rademacher complexity and $\mathbb{E}_\eta R_n\mathcal{G}$ is its expectation holding fixed the data, i.e. over the i.i.d. Rademacher variables η_i. The argument given in Section A.2.2 of Farrell, Liang, and Misra (2021) does not apply directly to $\mathbb{E}_\eta R_n\mathcal{G}$ because $\boldsymbol{\theta}$ is vector valued. Instead, we replace Lemma 2 therein with Maurer (2016, Corollary 1), which in our context yields (below η_{ik}'s denote i.i.d. Rademacher random variables)

$$
\mathbb{E}_\eta \sup_{\boldsymbol{\theta}\in\mathcal{F}^0_{\text{DNN}}}\frac{1}{n}\sum_{i=1}^n \eta_i\left(\ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}(\boldsymbol{x})) - \ell(\boldsymbol{y},\boldsymbol{t},\boldsymbol{\theta}_0(\boldsymbol{x}))\right) \leq \sqrt{2}C_\ell\mathbb{E}_\eta \sup_{\boldsymbol{\theta}\in\mathcal{F}^0_{\text{DNN}}}\sum_{k=1}^{d_{\boldsymbol{\theta}}}\frac{1}{n}\sum_{i=1}^n \eta_{ik}\big(\theta_k(\boldsymbol{x}_i) - \theta_{0k}(\boldsymbol{x}_i)\big)
$$

$$\leq \sqrt{2}C_\ell \sum_{k=1}^{d_{\boldsymbol{\theta}}} \mathbb{E}_\eta \sup_{\theta_k \in \mathcal{F}^0_{\text{DNN},k}} \frac{1}{n} \sum_{i=1}^n \eta_{ik} \big(\theta_k(\boldsymbol{x}_i) - \theta_{0k}(\boldsymbol{x}_i)\big),$$

with the second inequality following because the class of DNNs $\mathcal{F}_{\text{DNN}}$ we use is decomposable with respect to each coordinate, and therefore we can bound one coordinate at a time.

We then apply Section A.2.1 and Lemmas 3 and 4 of Farrell, Liang, and Misra (2021) to the term for each component function θ_k, $k = 1, \ldots, d_{\boldsymbol{\theta}}$, yielding

$$\mathbb{E}_\eta \sup_{\theta_k \in \mathcal{F}^0_{\text{DNN},k}} \frac{1}{n} \sum_{i=1}^n \eta_{ik} \big(\theta_k(\boldsymbol{x}_i) - \theta_{0k}(\boldsymbol{x}_i)\big) \leq 32 r_0 \sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{DNN},k})}{n} \left(\log \frac{2eM}{r_0} + \frac{3}{2} \log n\right)},$$

with probability $1 - \exp^{-\gamma}$, where $\text{Pdim}(\mathcal{F})$ is the pseudo-dimension of the class $\mathcal{F}$. Therefore, whenever $r_0 \geq 1/n$ and $n \geq (2eM)^2$,

$$\mathbb{E}_\eta \sup_{\boldsymbol{\theta} \in \mathcal{F}^0_{\text{DNN}}} \frac{1}{n} \sum_{i=1}^n \eta_i \left(\ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) - \ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}_0(\boldsymbol{x}))\right) \leq K r_0 \sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{DNN}})}{n} \log n},$$

for a constant K that depends on C_ℓ and $d_{\boldsymbol{\theta}}$.

This last bound is then combined with (C.2) and put into (C.1) and we find that

$$\begin{aligned}
&c_1 \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] \\
&\quad \leq 6K r_0 \sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{DNN}})}{n} \log n} + \sqrt{\frac{2C_\ell^2 r_0^2 \gamma}{n}} + \frac{23 \cdot 3MC_\ell}{3} \frac{\gamma}{n} + c_2 \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \gamma}{n}} + \frac{7C_\ell M \gamma}{n} \\
&\quad \leq r_0 \left(6K \sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{DNN}})}{n} \log n} + \sqrt{\frac{2C_\ell^2 \gamma}{n}}\right) + c_2 \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \gamma}{n}} + K_2 \frac{\gamma}{n} \\
&\quad \leq r_0 \left(K_1 \sqrt{\frac{WL \log(W)}{n} \log n} + \sqrt{\frac{2C_\ell^2 \gamma}{n}}\right) + c_2 \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \gamma}{n}} + K_2 \frac{\gamma}{n},
\end{aligned} \tag{C.3}$$

for constants K_1 and K_2, where the final inequality applies Theorem 6 in Bartlett et al. (2017) to bound the pseudo-dimension of ReLU networks in terms of their depth L and total parameters W.

The bound of Equation (C.3), reached under the assumption that $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2]^{1/2} \leq r_0$, provides the key input into Sections A.2.3 and A.2.4 of Farrell, Liang, and Misra (2021), which now go through with only change to the constants to capture the dependence on $d_{\boldsymbol{\theta}}$. Following those

steps exactly we find that with probability $1 - e^{-\gamma_1}$,

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] &\leq C\left(\frac{WL\log(W)}{n}\log n + \frac{\log\log n + \gamma_1}{n} + \epsilon_n^2\right) \\ \mathbb{E}_n\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] &\leq C'\left(\frac{WL\log(W)}{n}\log n + \frac{\log\log n + \gamma_1}{n} + \epsilon_n^2\right), \end{aligned} \tag{C.4}$$

for positive constants C and C' which do not depend on n but depend on the constants given in Assumption 1 and as well as the dimensionalities, including $d_{\boldsymbol{\theta}}$.

To specialize this result to the MLP case, for which $W \leq CH^2L$, we use the approximation result from Theorem 1 of Yarotsky (2017), or its restatement in Lemma 7 of Farrell, Liang, and Misra (2021). This result tell us that for each θ_{0k}, the following holds for H, L, and the approximation error ϵ_n:

$$\begin{aligned} H = H(\epsilon_n) &\leq W(\epsilon_n)L(\epsilon_n) \leq C^2\epsilon_n^{-\frac{d_c}{p}}(\log(1/\epsilon_n)+1)^2, \\ L = L(\epsilon_n) &\leq C\cdot(\log(1/\epsilon_n)+1). \end{aligned}$$

Therefore, a network that is $d_{\boldsymbol{\theta}}$ times wider can yield the same approximation for $\boldsymbol{\theta}_0$.

Importantly, only d_c matters here. To see why, suppose X_1 is binary. Then for two smooth, $d_{\boldsymbol{X}} - 1$-dimensional functions $\theta_{0k,1}$ and $\theta_{0k,0}$, it holds that $\theta_{0k}(\boldsymbol{x}) = x_1\theta_{0k,1}(x_2,\ldots,x_{d_{\boldsymbol{X}}}) + (1 - x_1)\theta_{0k,0}(x_2,\ldots,x_{d_{\boldsymbol{X}}})$. Adding a single node to each hidden layer allows the network to pass forward the input x_1 and multiply it with two separate learned functions just prior to the parameter, giving exactly $\bar{\theta}_{k,n}(\boldsymbol{x}) = x_1\bar{\theta}_{0k,1}(x_2,\ldots,x_{d_{\boldsymbol{X}}}) + (1-x_1)\bar{\theta}_{0k,0}(x_2,\ldots,x_{d_{\boldsymbol{X}}})$. (Intuitively, this is the same as the combination of $\widehat{\alpha}(\boldsymbol{x})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{x})$ in Figure B.1, with $\boldsymbol{t}$ there playing the role of the binary x_1 here.) The same argument can be applied to every category of the discrete data and to each function to be learned. Since $d_{\boldsymbol{X}}$ is fixed, this results in only a constant increase in the width of the network. Put together, we take $\epsilon_n = n^{-\frac{p}{2(p+d_c)}}$, i.e. $H \asymp n^{\frac{d_c}{2(p+d_c)}}\log^2 n$, $L \asymp \log n$, and we obtain the final result. □

C.2 Influence Function and Asymptotic Normality

To prove Theorem 2 we first derive the influence function, then we apply standard DML results to obtain the limiting distribution. Our derivation of the influence function applies Newey (1994). For

deeper treatments of influence functions, including efficiency bounds and conditions for existence, see Newey (1990), Newey (1994), van der Vaart (1998, Chapter 25) and Ichimura and Newey (2022).

The starting point is a parametric submodel, indexed by a parameter η. Because our first stage (3.3) is explicitly based on enriching the parametric structural model (3.1), our submodels are as well-behaved as the original model. For the calculation, distributions and other nonparametric objects are indexed by η, and thus we define $\boldsymbol{\theta}(\boldsymbol{x};\eta)$ and $\boldsymbol{\mu}_0(\eta)$ as

$$\boldsymbol{\theta}(\cdot;\eta) = \arg\min_{\boldsymbol{b}} \int \ell\left(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x})\right) f_{\boldsymbol{w}}(\boldsymbol{w};\eta) d\boldsymbol{w} \tag{C.5}$$

and

$$\boldsymbol{\mu}(\eta) = \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};\eta); \boldsymbol{t}^*\right) f_{\boldsymbol{x}}(\boldsymbol{x};\eta) d\boldsymbol{x}, \tag{C.6}$$

where $f_{\boldsymbol{w}}$ and $f_{\boldsymbol{x}}$ are the distributions of $\boldsymbol{w} = (\boldsymbol{y}', \boldsymbol{t}', \boldsymbol{x}')'$ and $\boldsymbol{x}$ respectively. For notational simplicity, we will assume throughout the derivation that such densities exist. The true data generating process is obtained at $\eta = 0$. When evaluating at $\eta = 0$ we will often omit the dependence on η, such as $f_{\boldsymbol{x}}(\boldsymbol{x};\eta) = f_{\boldsymbol{x}}(\boldsymbol{x})$, $\boldsymbol{\theta}(\boldsymbol{x};0) = \boldsymbol{\theta}_0(\boldsymbol{x})$, or $\mathbb{E}[\cdot]$ for expectations with respect to the true distribution.

The pathwise derivative approach proceeds, as in Newey (1994) and others, by finding a function $\boldsymbol{\psi}(\boldsymbol{w})$ such that

$$\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{W}) S(\boldsymbol{W})], \tag{C.7}$$

for the (true) score $S(\boldsymbol{w}) = S(\boldsymbol{w};\eta)|_{\eta=0}$.

The first step is differentiating (C.6) with respect to the parameter η, and evaluating this at $\eta = 0$. The product rule and the chain rule yield

$$\begin{aligned}
\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} &= \left.\frac{\partial}{\partial \eta}\left\{\int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};\eta); \boldsymbol{t}^*\right) f_{\boldsymbol{x}}(\boldsymbol{x};\eta) d\boldsymbol{x}\right\}\right|_{\eta=0} \\
&= \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};0); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} + \int \left.\frac{\partial \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};\eta); \boldsymbol{t}^*\right)}{\partial \eta}\right|_{\eta=0} f_{\boldsymbol{x}}(\boldsymbol{x};0) d\boldsymbol{x}, \\
&= \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} + \int \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*) \boldsymbol{\theta}_\eta(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x},
\end{aligned} \tag{C.8}$$

where $\boldsymbol{\theta}_\eta(\boldsymbol{x}) = \boldsymbol{\theta}_\eta(\boldsymbol{x};0)$ is the $d_{\boldsymbol{\theta}}$-vector gradient of $\boldsymbol{\theta}(\boldsymbol{x};\eta)$ with respect to η, evaluated at $\eta = 0$,

given by

$$\boldsymbol{\theta}_\eta(\boldsymbol{x};0) = \left.\frac{\partial \boldsymbol{\theta}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0},$$

and $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*)$ is the $d_{\boldsymbol{\mu}} \times d_{\boldsymbol{\theta}}$ Jacobian of $\boldsymbol{H}$ with respect to $\boldsymbol{\theta}$, evaluated at $\eta = 0$, that is, the matrix with $\{h,k\}$ element, for $h = 1,\ldots,d_{\boldsymbol{\mu}}, k = 1,\ldots,d_{\boldsymbol{\theta}}$, given by

$$\left[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*)\right]_{h,k} = \left.\frac{\partial H_h(\boldsymbol{x},\boldsymbol{b};\boldsymbol{t}^*)}{\partial b_k}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};0)},$$

with H_h the h^{th} element of $\boldsymbol{H}$ and b_k the k element of $\boldsymbol{b}$. For intuition, note that element $h = 1,\ldots,d_{\boldsymbol{\mu}}$ of the $d_{\boldsymbol{\mu}}$-vector $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*)\boldsymbol{\theta}_\eta(\boldsymbol{x})$ is

$$\left.\frac{\partial \boldsymbol{H}_h}{\partial \eta}\right|_{\eta=0} = \sum_{k=1}^{d_{\boldsymbol{\theta}}} \left.\frac{\partial \boldsymbol{H}_h(\boldsymbol{x},\boldsymbol{b};\boldsymbol{t}^*)}{\partial \boldsymbol{b}_k}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};0)} \left.\frac{\partial \boldsymbol{\theta}_k(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0}.$$

We will show that both terms of Equation (C.8) above can be written as expectations of products with the full score $S(\boldsymbol{w})$, as required by (C.7). We will often use the standard facts that scores are mean zero and that

$$S(\boldsymbol{y},\boldsymbol{x},\boldsymbol{t}) = S(\boldsymbol{y},\boldsymbol{t} \mid \boldsymbol{x}) + S(\boldsymbol{x}). \tag{C.9}$$

The first term of Equation (C.8) is

$$\begin{aligned}\int \boldsymbol{H}\left(\boldsymbol{x},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} &= \mathbb{E}\left[\boldsymbol{H}\left(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*\right) S(\boldsymbol{X})\right] \\ &= \mathbb{E}\left[\boldsymbol{H}\left(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*\right) S(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{T})\right], \end{aligned} \tag{C.10}$$

where the first equality holds because the marginal score obeys $S(\boldsymbol{x})f_{\boldsymbol{x}}(\boldsymbol{x}) = \partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)/\partial\eta|_{\eta=0}$ and the second equality follows from the usual mean zero property of scores and (C.9):

$$\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{x}),\boldsymbol{t}^*)S(\boldsymbol{Y},\boldsymbol{T} \mid \boldsymbol{X})\right] = \mathbb{E}\Big[\boldsymbol{H}(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{x}),\boldsymbol{t}^*)\mathbb{E}\left[S(\boldsymbol{Y},\boldsymbol{T} \mid \boldsymbol{X}) \mid \boldsymbol{X}\right]\Big] = 0.$$

This first term is then the standard "plug-in" portion of the influence function, that is, the term that would appear if $\boldsymbol{\theta}_0(\boldsymbol{x})$ were known (or if $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ were fixed). The second term of Equation (C.8) will give rise to the correction factor that accounts for the nonparametric estimation.

To find this correction factor, we must find $\boldsymbol{\theta}_\eta(\boldsymbol{x}) = \partial\boldsymbol{\theta}(\boldsymbol{x};\eta)/\partial\eta|_{\eta=0}$. This is a key step in the derivation, and crucially leverages the structure of the model ℓ and the fact that ℓ depends on $\boldsymbol{\theta}(\cdot)$ only through evaluation at a single point and only through $\boldsymbol{X}$. We will use these facts to derive and expression for $\partial\boldsymbol{\theta}(\boldsymbol{x};\eta)/\partial\eta$, which involves the appropriate scores and then may be substituted into (C.8) to yield the required form.

We begin with the fact that the first order condition holds as an identity in η and conditional on $\boldsymbol{X}$. That is, as an identity in η,

$$\mathbb{E}_\eta\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W},\boldsymbol{\theta}(\boldsymbol{x};\eta))\middle|\,\boldsymbol{X}=\boldsymbol{x}\right]\equiv 0, \tag{C.11}$$

where $\boldsymbol{\ell}_{\boldsymbol{\theta}}$ is the $d_{\boldsymbol{\theta}}$-vector gradient of ℓ with respect to $\boldsymbol{\theta}$, given by

$$\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}(\boldsymbol{x};\eta)) = \left.\frac{\partial\ell\left(\boldsymbol{w},\boldsymbol{b}\right)}{\partial\boldsymbol{b}}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};\eta)}.$$

The expectation is also indexed by η in the submodel, as the density depends on η. To be explicit, as an identity in η we have

$$\int \left.\frac{\partial\ell\left(\boldsymbol{w},\boldsymbol{b}\right)}{\partial\boldsymbol{b}}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};\eta)} f_{\boldsymbol{y},\boldsymbol{t}|\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{t};\eta\mid\boldsymbol{x})d\boldsymbol{y}d\boldsymbol{t}\equiv 0.$$

Define $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}(\boldsymbol{x};\eta))$ as the $d_{\boldsymbol{\theta}}\times d_{\boldsymbol{\theta}}$ matrix of second derivatives of $\ell\left(\boldsymbol{w},\boldsymbol{b}\right)$ with respect to $\boldsymbol{b}$, evaluated at $\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};\eta)$. That is, $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}(\boldsymbol{x};\eta))$ has $\{k_1,k_2\}$ element given by

$$\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}(\boldsymbol{x};\eta))\right]_{k_1,k_2} = \left.\frac{\partial^2\ell\left(\boldsymbol{w},\boldsymbol{b}\right)}{\partial b_{k_1}\partial b_{k_2}}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};\eta)},$$

where b_{k_1} and b_{k_2} are the respective elements of $\boldsymbol{b}$. With this notation, differentiating the above identity with respect to η and applying the chain rule we find

$$\int \left.\frac{\partial\ell\left(\boldsymbol{w},\boldsymbol{b}(\boldsymbol{x})\right)}{\partial\boldsymbol{b}}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x};\eta)} \frac{\partial f_{\boldsymbol{y},\boldsymbol{t}|\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{t};\eta\mid\boldsymbol{x})}{\partial\eta}d\boldsymbol{y}d\boldsymbol{t} + \int \boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}(\boldsymbol{x};\eta))\boldsymbol{\theta}_\eta(\boldsymbol{x};\eta)f_{\boldsymbol{y},\boldsymbol{t}|\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{t};\eta\mid\boldsymbol{x})d\boldsymbol{y}d\boldsymbol{t}=0,$$

where the second term captures the derivatives of $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}(\boldsymbol{x};\eta))$ with respect to η, and recall, $\boldsymbol{\theta}_\eta(\boldsymbol{x};\eta)$

is the $d_{\boldsymbol{\theta}}$-vector gradient of $\boldsymbol{\theta}$ with respect to η, and is the key ingredient.

Evaluating this result at $\eta = 0$, we obtain

$$\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}) \middle| \boldsymbol{X}\right] + \mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))\boldsymbol{\theta}_\eta(\boldsymbol{x}) \middle| \boldsymbol{X}\right] = 0, \tag{C.12}$$

where $S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X})$ is the conditional score and is obtained because $S(\boldsymbol{y}, \boldsymbol{t} \mid \boldsymbol{x}) f_{\boldsymbol{y},\boldsymbol{t}|\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{t} \mid \boldsymbol{x}) = \partial f_{\boldsymbol{y},\boldsymbol{t}|\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{t}; \eta \mid \boldsymbol{x})/\partial\eta\big|_{\eta=0}$. Rearranging (C.12), and using that $\boldsymbol{\theta}$ is only a function of $\boldsymbol{X}$, gives

$$\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x})) \middle| \boldsymbol{X}\right] \boldsymbol{\theta}_\eta(\boldsymbol{x}) = -\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}) \middle| \boldsymbol{X}\right].$$

Then, because $\boldsymbol{\Lambda}(\boldsymbol{x}) := \mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x})) \middle| \boldsymbol{X} = \boldsymbol{x}\right]$ is invertible, we have

$$\begin{aligned}
\boldsymbol{\theta}_\eta(\boldsymbol{x}) &= -\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x})) \middle| \boldsymbol{X}\right]^{-1} \mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}) \middle| \boldsymbol{X}\right] \\
&= -\mathbb{E}\left[\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}) \middle| \boldsymbol{X}\right].
\end{aligned}$$

Substituting this into the second term of Equation (C.8) and applying iterated expectations, we have

$$\begin{aligned}
\int \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*)\boldsymbol{\theta}_\eta(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} &= -\mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\mathbb{E}\left[\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}) \middle| \boldsymbol{X}\right]\Big] \\
&= -\mathbb{E}\Big[\mathbb{E}\left[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X}) \middle| \boldsymbol{X}\right]\Big] \\
&= -\mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T} \mid \boldsymbol{X})\Big].
\end{aligned}$$

Next, because the first order condition holds conditionally,

$$\begin{aligned}
&\mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{X})\Big] \\
&\qquad = \mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x})) \mid \boldsymbol{X}\right] S(\boldsymbol{X})\Big].
\end{aligned}$$

Therefore, continuing from the previous display and applying (C.9), the second term of Equation (C.8) is of the required form:

$$-\mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{X})\Big] \tag{C.13}$$

Combining Equations (C.10) and (C.13) with (C.8), we find that

$$\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} = \mathbb{E}\left[\boldsymbol{H}\left(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) S(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{T})\right] - \mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W}, \boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{X})\Big]. \quad \text{(C.14)}$$

Thus we have verified Equation (C.7) with

$$\boldsymbol{\psi}(\boldsymbol{w}) = \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) - \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}_0(\boldsymbol{x})). \quad \text{(C.15)}$$

This is not an influence function as it lacks the appropriate centering, but of course $\mathbb{E}[\boldsymbol{\mu}_0 S(\boldsymbol{W})] = \boldsymbol{\mu}_0\mathbb{E}[S(\boldsymbol{W})] = 0$, and thus we can freely center this $\boldsymbol{\psi}(\boldsymbol{t})$ and still obey (C.7). □

C.2.1 Asymptotic Normality

The asumptotic Normality of Theorem 2 follows from Theorems 3.1 and 3.2 of Chernozhukov et al. (2018a) upon verifying Assumptions 3.1 and 3.2 therein. Assumption 3.1(a) holds for $\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_0$ given in Equation 4.1: the first term of $\boldsymbol{\psi}$ has mean $\boldsymbol{\mu}_0$ by definition in (3.4) while the second is (conditionally) mean zero as assumed in Assumption 3, with $\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}$ uniformly bounded. Assumption 3.1(b), linearity, holds by definition of (3.4) and the form of the score in Equation 4.1. Assumption 3.1(c) holds by Assumption 3, in particular the assumed smoothness and the nonsingularity of $\boldsymbol{\Lambda}(\boldsymbol{x})$. Assumption 3.1(d), Neyman orthogonality, is verified directly by the calculation of the influence function. Assumption 3.1(e) holds trivially as the matrix J_0 therein is the identity. Assumption 3.2, parts (b) and (d) follow directly from the moment conditions imposed. Conditions (a) and (c) follow from Equations (3.7) and (3.8) of Chernozhukov et al. (2018a) and the assumed convergence of the first stage estimates. □

C.3 Proof of Corollary B.1

For this derivation, define $\widetilde{\boldsymbol{T}} = (1, \boldsymbol{T}')'$ and recall that $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0')'$ (and similarly for realizations, estimators, etc). First, consider identification. Since G is invertible and conditional expectations are always identified, the quantity $G^{-1}(\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}])$ is identified. Suppose that $\boldsymbol{\theta}_0(\boldsymbol{x})$ is not identified. Then there exists $\boldsymbol{\theta}_1(\boldsymbol{x})$ and $\boldsymbol{\theta}_2(\boldsymbol{x})$ such that $G^{-1}\left(\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}]\right) = \boldsymbol{\theta}_1(\boldsymbol{x})'\widetilde{\boldsymbol{t}} =$

$\boldsymbol{\theta}_2(\boldsymbol{x})'\widetilde{\boldsymbol{t}}$ a.e. or equivalently that for $\boldsymbol{\theta}_*(\boldsymbol{x}) = \boldsymbol{\theta}_1(\boldsymbol{x}) - \boldsymbol{\theta}_2(\boldsymbol{x})$, $\boldsymbol{\theta}_*(\boldsymbol{x})'\widetilde{\boldsymbol{t}} = 0$. But $\boldsymbol{\theta}_*(\boldsymbol{x})'\widetilde{\boldsymbol{t}} = 0$ a.e. implies that

$$0 = \mathbb{E}\left[\left(\boldsymbol{\theta}_*(\boldsymbol{x})'\widetilde{\boldsymbol{t}}\right)^2 \mid \boldsymbol{X} = \boldsymbol{x}\right] = \boldsymbol{\theta}_*(\boldsymbol{x})'\mathbb{E}[\widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{T}}'|X]\boldsymbol{\theta}_*(\boldsymbol{x}),$$

but because the middle matrix is positive definite, this means that $\boldsymbol{\theta}_*(\boldsymbol{x})$ is zero. For linear G, this argument is given in Huang and Shen (2004), among others.

The estimation bounds follow immediately from Theorem 1, given the conditions of Assumption 4. The fact that $\mathbb{E}[\widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{T}}' \mid \boldsymbol{X}]$ is (uniformly) positive yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{X})'\widetilde{\boldsymbol{T}} - \boldsymbol{\theta}_0(\boldsymbol{X})'\widetilde{\boldsymbol{T}}\right)^2\right] &= \mathbb{E}_{\boldsymbol{X}}\left[\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\right)'\mathbb{E}[\widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{T}}' \mid \boldsymbol{X}]\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\right)\right] \\ &\geq C\mathbb{E}_{\boldsymbol{X}}\left[\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\right)'\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\right)\right].\end{aligned}$$

This verifies the curvature condition on the loss function. The continuity condition holds because the loss is smooth in g and the linear index can be recovered from $g(G(\boldsymbol{\theta}(\boldsymbol{x})'\widetilde{\boldsymbol{t}}))$. The structure of the network ensures that the network and the smoothness of the loss imply that the approximation and bounds immediately apply to the function $g(G(\boldsymbol{\theta}(\boldsymbol{x})'\widetilde{\boldsymbol{t}}))$, and the smoothness of these functions mean that the linear index $\boldsymbol{\theta}(\boldsymbol{x})'\widetilde{\boldsymbol{t}}$ can be recovered. □