arXiv:2010.14694v2 [econ.EM] 23 Jul 2021

Deep Learning for Individual Heterogeneity: An Automatic Inference Framework*

Max H. Farrell Tengyuan Liang Sanjog Misra

University of Chicago, Booth School of Business

July 27, 2021

Abstract

We develop methodology for estimation and inference using machine learning to enrich economic models. Our framework takes a standard economic model and recasts the parameters as fully flexible nonparametric functions, to capture the rich heterogeneity based on potentially high dimensional or complex observable characteristics. These "parameter functions" retain the interpretability, economic meaning, and discipline of classical parameters. In contrast to common implementations of machine learning in economics, these functions need not be predictions. We show that deep learning is particularly well-suited to structured modeling of heterogeneity in economics. First, we show how the network architecture can be easily designed to match the global structure of the economic model, delivering novel methodology that moves deep learning beyond prediction. Second, we prove convergence rates for the estimated parameter functions. These parameter functions are then the key input into the finite-dimensional parameter of inferential interest. We obtain valid inference based on a novel orthogonal score or influence function calculation that covers any second-stage parameter and any machine-learning-enriched model that uses a smooth per-observation loss function. No additional derivations are required and the score can be taken directly to data, using automatic differentiation if needed to obtain the components: the researcher need only define the original model and define the parameter of interest. A key insight is that we need not write down the influence function in order to evaluate it on the data. We apply this after deep learning, but our result can be used for any first-step estimator. Our framework covers, as special cases, well-known examples such as average treatment effects and partially linear models, but we also seamlessly deliver new results for such diverse examples as price elasticities, willingness-to-pay, and surplus measures in binary or multinomial choice models, average marginal and partial effects of continuous treatment variables, fractional outcome models, count data, heterogeneous production function components, and more. Across all these contexts inference can be made as automated as is currently available in special cases. We illustrate the utility of our framework with an application to a large scale advertising experiment for short-term loans. We show how economically meaningful estimates and inferences can be made that would be unavailable without our framework.

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

*The authors would like to thank Chris Hansen and Whitney K. Newey, the participants and discussants at the Chamberlain seminar and 2020 QME conference, as seminar participants at Columbia, NYU, UC Berkeley, UC Santa Barbara for useful discussions, comments, and suggestions.

1 Introduction

The goal of this paper is to leverage modern machine learning and rich data to capture individual heterogeneity in the context of economic models. Parametric, structural models are a cornerstone of applied research in economics and social sciences. The parameters estimated in these models are interpretable, useful for policy, and disciplined by economic principles. We develop a methodology to maintain these advantages while simultaneously incorporating machine learning methods for flexibly estimating heterogeneity. Our idea is to recast the parameters of a model as flexible functions themselves: enriching the model without losing the structure. We estimate the parameters using deep learning, for which we provide new results. We then deliver second-step inference by deriving a novel influence function that applies to any such enriched model.

The starting point is a researcher-specified model that relates outcomes $\boldsymbol{Y}$ to the covariates of central interest, the policy or treatment variables $\boldsymbol{T}$. This model is encapsulated by a loss function $\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta})$ for a vector of parameters $\boldsymbol{\theta}$ that are estimated from the data. The model encodes structure that is grounded in economic principles and economic reasoning. To fix ideas, take the context of our empirical application where we revisit the data of Bertrand et al. (2010). Here $\boldsymbol{Y}$ is a customer's binary choice to apply for a loan and $\boldsymbol{T}$ are characteristics of the loan and an advertisement received, one of which is the interest rate offered. A standard approach to this problem would be a (structural) logistic binary choice model where the parameters $\boldsymbol{\theta}$ are the coefficients on $\boldsymbol{T}$, including an intercept. Such a model has numerous advantages. First, the parameters have a clear and direct interpretation, and generally respect economic theory. Second, economically meaningful, and policy-relevant, summary parameters are easily computed, such as elasticities or measures of surplus. Further, the economic structure can be used to answer substantial policy questions. For example, although interest rates are only observed at certain values, we can use the model to study what would occur at other levels. Indeed, from basic economic principles like profit maximization, we can compute the optimal interest rate as a function of the parameters, say $r^*(\boldsymbol{\theta})$. All of these are only possible because of the economic structure imposed on the analysis.

However, if there is heterogeneity, which is almost a given in most contexts, the parameter estimates may not be reliable in practice, and this has spurred a push to move beyond rigid parametric models, which can only crudely capture heterogeneity, if at all. Heterogeneity can come in many

forms or depend on many things. This fact, along with the increasing availability of large, complex data sets, has motivated the adoption of novel machine learning methods in economics, allowing researchers to study economic phenomena at levels of detail previously not possible.

Our approach to this problem allows for the use of powerful machine learning to capture rich heterogeneity, while simultaneously maintaining the structure and interpretability of the economic model, with all its advantages. To do this, we recast the model's parameters as functions of observed characteristics $\boldsymbol{X}$, thus enriching the model to $\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X}))$. These "parameter functions" allow for fully flexible heterogeneity but keep intact the structure of the economic model connecting $\boldsymbol{Y}$ and $\boldsymbol{T}$. In general we will not know either the functional form of $\boldsymbol{\theta}_0(\boldsymbol{x})$ nor which covariates are important. This is one strong motivation for applying modern machine learning methods. Our approach exploits the flexibility of machine learning within the structure dictated by economic models.

We thus deploy machine learning to directly estimate meaningful objects, which has several advantages compared to using ML to only obtain predictions. First, for any set of characteristics $\boldsymbol{X} = \boldsymbol{x}$, $\boldsymbol{\theta}(\boldsymbol{x})$ gives the effect for an individual of "type" $\boldsymbol{x}$, and therefore retains all of the meaning, interpretability and usefulness of the original parameters. Additionally, we can "score" or "type" future individuals through their characteristics, because $\boldsymbol{\theta}(\boldsymbol{x})$ captures heterogeneity that can be used for future policy. Second, because we have maintained the economic model, we can leverage its structure to answer substantive questions. For example, in our data we can compute the personalized, targeted optimal interest rate $r^*(\boldsymbol{\theta}(\boldsymbol{x}))$.

Our approach of implementing ML by enriching an economic model thus directly address several major drawbacks of common applications of ML in economics. Typically, ML have largely been confined to prediction problems, or those that can be convert to prediction problems. These prediction functions are added to models only as nuisance functions. The term "nuisance" here is illustrative: it connotes that flexibility is required in some piece of the model, but is not per se interesting or meaningful. This is typical not just of ML, but of semiparametrics more broadly, and we aim to depart from this mindset. In the context of our application, for example, this might mean using ML to classify potential borrowers with an unstructured function $\mathbb{E}[Y|\boldsymbol{t}, \boldsymbol{x}] = \theta(\boldsymbol{T}, \boldsymbol{X})$: a pure prediction problem. Such an estimate would not only have worse statistical properties, it would lose the economic meaning of the original model. The results would not be interpretable directly and useful quantities would be difficult or impossible to extract. Further, without the discipline of the

economic principles of the model, such estimates may make little sense.

To implement our approach we require estimates of the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ and an inference engine for second-stage parameters of interest. We gives results for both steps. In both cases we make heavy use of the idea that we have enriched an economic model: it is this concrete structure that enables us to deliver broad and powerful results. For first stage estimation, we show that deep neural networks (DNNs) have a unique combination of strengths which makes them able to directly incorporate the structure of the economic model as well as handle modern data sets and complex heterogeneity. We prove new convergence rates in this context. For second step inference we give a new orthogonal score that can be widely and easily applied.

Deep neural networks have had incredible empirical success, matching or setting the state of the art in a wide range of tasks. But this has largely been in prediction problems, and indeed, most software is designed only to optimize prediction loss functions. We develop a novel, yet simple, architectural idea so that the global structure of the economic model is baked into the estimation directly and the parameter functions are recovered, instead of focusing on prediction. The idea is simple and implementation is straightforward, but this appears to have been mostly overlooked in the ML literature. Our ideas are in the vein of nonparametric M-estimation, which has a longer history but has relied on classical methods which are not equipped to handle the complex heterogeneity that is the hallmark of modern applications. DNNs also handle discrete data seamlessly in practice (as well as in theory), in contrast to classical methods. We prove new results for structured deep neural network estimates of $\boldsymbol{\theta}(\boldsymbol{X})$, which crucially recover the parameter functions, not predictions, and depend on the dimension of $\boldsymbol{X}$, the heterogeneity, independent of the dimension of the variables of interest $\boldsymbol{T}$. Our results build on the recent work of Farrell et al. (2021) and contribute more broadly to the nonparametric M-estimation literature (see Chen (2007) for review and references).

Next, a feasible inference engine is required following the deep learning of $\boldsymbol{\theta}(\boldsymbol{X})$ in the enriched model. We obtain valid inference on a finite-dimensional parameter of interest $\boldsymbol{\mu}$ that depends directly on the parameter functions $\boldsymbol{\theta}(\boldsymbol{X})$. We achieve this by characterizing an influence function, or orthogonal score, for $\boldsymbol{\mu}$, and then making use of the recent results for influence function based estimators following machine learning. Our derivation builds directly on long-standing ideas in semiparametrics, chiefly Newey (1994), and for inference we follow the method of Chernozhukov et al. (2018), combining an orthogonal score with sample splitting. A drawback of this general approach

to semiparametric inference is that the influence function must be known in advance, and this has perhaps hampered take-up among applied researchers. Most applications focus on a few special cases where the influence function is known (e.g. the case of average treatment effects). Our contribution to inference is to derive a single influence function that includes the correction term for any smooth per-observation loss $\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X}))$. Thus, for any ML-enriched economic model inference can proceed without further calculations. We apply this after deep learning, but our influence function can be used in general, for any first step estimator meeting standard conditions, such as lasso, trees and forests, or classical sieve or kernel methods. Our influence functions recovers well known special cases such as average treatment effects and partially linear models, but immediately delivers new results for many other contexts, including selection models, choice models, fractional outcomes, and more broadly, smooth QMLE contexts and other such areas.

A key insight is that we only need to characterize the influence function and evaluate its empirical analogue at each data point, consequently obtaining a point estimator and standard errors, without needing to explicitly write it down. This can be contrasted with the typical approach of writing down the influence function, or orthogonal score, explicitly and then plugging in estimates of each nuisance function. Our goal is to make inference feasible in a wide variety of settings, not to focus on the properties of the score, such as efficiency comparisons, and this shift of mindset allows for broadly applicable methodology. An important tool here is automatic differentiation. The influence function depends only upon *ordinary* derivatives of the model ℓ and of the function defining the parameter of interest, as though the model were still parametric with homogeneous effects. In many cases, the derivatives of the model are already well known, and these forms can be used, but if not, the derivatives can be evaluated on the data automatically. For example, recall that in our empirical study, $r^*(\boldsymbol{\theta}(\boldsymbol{x}))$ is the personalized optimal interest rate. This $r^*(\boldsymbol{\theta}(\boldsymbol{x}))$ is not available in closed form, but rather as the solution to a fixed point problem. Therefore, the influence function for a parameter such as expected profits at the optimal personalized interest rates $\mu = \mathbb{E}[\pi(r^*(\boldsymbol{\theta}(\boldsymbol{X})))]$ cannot be explicitly written down. Nonetheless, because μ is a smooth function of $r^*(\boldsymbol{\theta})$, and $r^*(\boldsymbol{\theta})$ depends smoothly on $\boldsymbol{\theta}$, its derivatives can be evaluated at each data point with no trouble, allowing feasible inference. This would not be possible without our approach that retains the economic structure of the model.

Taken together, the combination of the specification we adopt, the availability of computing

infrastructure, and the theory we present, offers a perfect package for applied researchers across economics and social science hoping to exploit ML but maintain discipline-specific knowledge and interpretability. Our work should be broadly useful by delivering a tractable and valid estimation and inference framework that covers many interesting contexts. The next section presents an overview of our methodological framework and its interpretation, and gives a brief overview of the main results. Our results related to many strands of recent literature, and we discuss these in context as they arise. Section 3 shows how deep learning is an excellent tool in our context and gives theoretical results. Section 4 discusses semiparametric inference, our novel influence function, and asymptotic normality. We apply our methods in Section 5 to an empirical study of short term loan applications. Section 6 gives a sense of the breadth of applicability of our results by discussing a number of examples, but is by no means exhaustive. Extensions are discussed in Section 7 and finally Section 8 concludes. Proofs are given in the appendix.

2 A Methodological Framework for Enriching Economic Models with Machine Learning

In this section we describe our framework enriching economic models to capture individual heterogeneity and give an informal summary of our results. The starting point is a standard, parametric economic model. We assume the researcher observes data on outcomes, $\boldsymbol{Y} \in \mathbb{R}^{d_{\boldsymbol{Y}}}$, and on the covariates of central interest, or treatments, $\boldsymbol{T} \in \mathbb{R}^{d_{\boldsymbol{T}}}$, which can be continuous, discrete, or mixed.[1] The researcher relates these two with an economic model, which is encapsulated by a parametric per-observation loss function, $\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta})$, indexed by a parameter $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$. For some parameter space $\boldsymbol{\Theta}$, dictated by the structure of the problem, the researcher then solves $\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}[\ell(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta})]$, that is, standard M-estimation, including all (psuedo-/quasi-) likelihood-like settings. Two key aspects of this approach are: (i) $\boldsymbol{\theta}$ are parameters, not predictions, and have economic meaning and (ii) the effects are homogeneous. We will retain (i) while removing (ii).

Our framework starts with the same parametric model, but *recasts* the parameters $\boldsymbol{\theta}$ as *functions* of observed individual characteristics $\boldsymbol{X} \in \mathbb{R}^{d_{\boldsymbol{X}}}$ to allow for heterogeneity. Thus, in place of $\boldsymbol{\theta}$ we

[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 $\|g\|_{L_2(\boldsymbol{X})} = \mathbb{E}[g(\boldsymbol{X})^2]^{1/2}$.

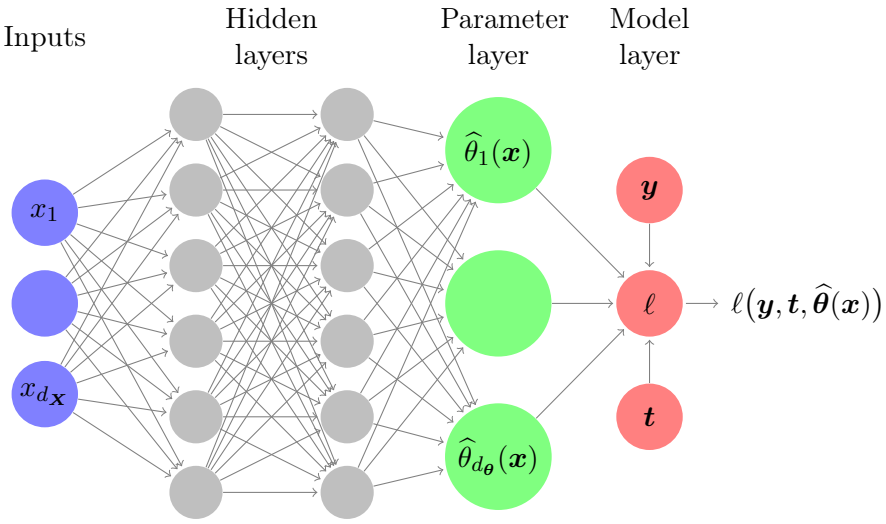


Figure 1: Illustration of the deep neural network estimation of the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ for a generic structured model (2.1)

will have $\boldsymbol{\theta}(\boldsymbol{x})$, mapping $\mathbb{R}^{d_{\boldsymbol{X}}} \mapsto \mathbb{R}^{d_{\boldsymbol{\theta}}}$, and we assume that the true parameter functions $\boldsymbol{\theta}_0(\cdot)$ solve

$$\boldsymbol{\theta}_0(\cdot) = \underset{\boldsymbol{\theta} \in \mathcal{H}}{\arg\min}\, \mathbb{E}\left[\ell\big(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X})\big)\right], \tag{2.1}$$

for an appropriate function class $\mathcal{H}$ (formalities are given below). We can thus capture heterogeneity in a fully flexible way, while retaining all the structure and interpretability of the standard model. For intuition, it is often useful to remember that at any value $\boldsymbol{X} = \boldsymbol{x}$, $\boldsymbol{\theta}(\boldsymbol{x})$ has the same interpretation as $\boldsymbol{\theta}$, but for the "type" determined by $\boldsymbol{x}$. It is also useful to think of individual-specific effects, which may be what researchers are implicitly worried about when heterogeneity is a concern, that is $\boldsymbol{\theta}_i$ minimizes $\ell(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{\theta})$ for each i. Such individual specific parameters are also as interpretable and meaningful as the original, homogeneous case, but of course $\boldsymbol{\theta}_i$ cannot typically be recovered from the data, and may not be useful in the future, as person i will not be seen again. One can think of $\boldsymbol{\theta}(\boldsymbol{x}_i)$ as an approximation to $\boldsymbol{\theta}_i$, one that uses all available information, and thus captures the portion of heterogeneity useful for future policy targeting.

The functions $\boldsymbol{\theta}(\boldsymbol{x})$ are not prediction functions necessarily, nor do we view them as nuisance functions, a term which implies they are uninteresting. Section 3 shows why deep learning is well suited for structured modeling and gives a novel, yet simple architectural idea for doing so. This means that we use deep learning to recover meaningful functions, thus moving ML away from

prediction tasks and toward estimation of scientifically and economically interesting objectives, a shift in mindset from the typical applications of machine learning. This implementation innovation and the convergence rates for the resulting estimated parameter functions are the two main contributions of this paper to deep learning, and to machine learning first stage parameters more broadly.

The key result is that we estimate the parameter functions at a fast enough rate for inference, and importantly, that the rate depends only upon the number of continuous heterogeneity covariates, denoted $d_C \leq d_{\boldsymbol{X}}$, and does not depend on dimension of the policy/treatment variables. That is, for our estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0(\boldsymbol{x})$ defined by (2.1), Theorem 1 establishes that

$$\left\|\widehat{\theta}_k - \theta_{0k}\right\|^2_{L_2(\boldsymbol{X})} = O(n^{-\frac{p}{p+d_C}} \log^8(n)), \qquad k = 1, \ldots, d_{\boldsymbol{\theta}},$$

provided ℓ is sufficiently smooth and curved near the truth. This result relies on our novel architectural idea, where the structure of the model is baked directly into the deep net architecture. The architecture, and hence the optimization of the loss, targets the parameter functions, not predictions. This idea is illustrated in Figure 1.

The heterogeneous parameter functions $\boldsymbol{\theta}_0(\boldsymbol{x})$ are then key inputs into the final parameter of inferential interest, denoted $\boldsymbol{\mu}_0 \in \mathbb{R}^{d_{\boldsymbol{\mu}}}$. For a known, smooth function $H : \{\mathbb{R}^{d_{\boldsymbol{X}}} \times \mathbb{R}^{d_{\boldsymbol{\theta}}}\} \mapsto \mathbb{R}^{d_{\boldsymbol{\mu}}}$, chosen by the researcher, we conduct inference on

$$\boldsymbol{\mu}_0 = \mathbb{E}\Big[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)\Big], \tag{2.2}$$

where $\boldsymbol{t}^*$ is some fixed value of interest. Many economically interesting statistics take this form, in particular, depending on functions that are not predictions. To make inference feasible we derive an influence function for any such $\boldsymbol{\mu}_0$, which includes deriving the correction factor for any ML-enriched M-estimation problem. Regularity conditions are below; in particular, we assume the $\boldsymbol{\mu}_0$ is pathwise differentiable. Beyond this, the form of $\boldsymbol{\mu}_0$ can be generalized at the cost of notation.

The main theoretical contribution to second stage inference is the influence function calculation, yielding a Neyman orthogonal score, that is specific enough to be directly implemented while still being general enough to cover any enriched structural model based on a smooth per-observation loss function. From a practical point of view, two key ideas in our work are the use of computational

differentiation (automatic or numerical) and the conceptual point of evaluating influence functions on the observed data rather than writing them down.

Theorem 2, in Section 4, gives the Neyman orthogonal score for any (sufficiently regular) such parameter $\boldsymbol{\mu}_0$ and first stage $\boldsymbol{\theta}(\boldsymbol{x})$ coming from an ML-enriched model. Importantly, this score depends only on *ordinary* derivatives. Let $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)$ and $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ be the gradients of H and ℓ with respect to $\boldsymbol{\theta}$ and denote $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}]$ the conditional expectation of the Hessian of ℓ, all evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x})$. Then the Neyman orthogonal score is $\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_0$, where

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

This can be taken to the data directly. That is, given estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Lambda}}$, we can evaluate $\boldsymbol{\psi}(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$ at every data point without further derivation, which is all that is required for feasible estimation and inference. For many standard models, these derivatives are known. If not, they can obtained using automatic differentiation tools or other computational methods. In other words, once the researcher specifies their economic model via $\ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta})$ and parameter of interest via $\boldsymbol{H}$, the full influence function is known and ready to use. This holds for any sufficiently smooth functions, even if not available in closed form, as in our example with the optimal interest rate and corresponding profits.

The general form of this orthogonal score and what each piece represents conceptually should call to mind the parametric case. If $\boldsymbol{\theta}_0$ were constant, then classical results for two-step estimation, as in Newey and McFadden (1994, Section 6), would yield the effect of the first stage on the second, and deliver an influence function that looks the same, but with $\boldsymbol{H}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Lambda}$ as constants, instead of conditional on $\boldsymbol{x}$. Thus our influence function result can be viewed as establishing the analogous result for fully nonparametric parameter functions. We view this familiarity as a strength, as it perfectly matches our core idea of enriching a well-understood economic model, and may help to demystify semiparametric inference.

Armed with this orthogonal score, we can obtain a point estimator $\widehat{\boldsymbol{\mu}}$ and standard errors $\widehat{\boldsymbol{\Psi}}$ such that $\widehat{\boldsymbol{\mu}} \overset{d}{\sim} \mathcal{N}(\mu_0, \widehat{\boldsymbol{\Psi}}/n)$, allowing inference on any aspect of the vector $\boldsymbol{\mu}_0$. 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 inference in practice is no more difficult than what is currently available for simple cases, like average treatment effects: we estimate the two nonparametric objects, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Lambda}}$, and plug them in as needed. Estimation of $\boldsymbol{\Lambda}(\boldsymbol{x})$ is discussed in more detail below, and can rely on standard ML methods, including deep learning, and may require sample splitting. However, an important methodological point is that when $\boldsymbol{T}$ is randomized this matrix can often be *computed*, as opposed to estimated. In general, two-step semiparametric inference often involves a denominator term such as this, and computing this term can yield more stable and robust results compared estimating it. See Remark 4 and our application in Section 5.

A special case of our framework that is useful for illustrating the main ideas, as well as prevalent in empirical and theoretical work, is when $\boldsymbol{Y}$ is a scalar Y and (2.1) is built around conditional mean restriction and the parameters are the intercept and the slopes. That is, let $\boldsymbol{t}$ include a constant term and assume that for a known function $G(u), u \in \mathbb{R}$,

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G\left(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}\right). \tag{2.3}$$

Recall that the $\boldsymbol{X}$ are observed; this is not a random coefficient model.[2] Rather, we have made the intercept and slope fully heterogeneous in observables. Clearly, (2.3) can be implemented using (2.1), given an appropriate loss function. A crucial piece will be the first order condition of that loss, that is, what orthogonality condition to use. For example, if G is the logistic link we may take ℓ to be the nonlinear least squares or the likelihood, which have different first order conditions, and this will change the influence function given later and thus impact implementation. O'Hagan (1978) may be the earliest treatment of a model like (2.3), and the structure here is often known as a "varying coefficient" model (Cleveland et al., 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). Our results speak directly to this

[2] One can consider the random coefficients model an alternative parametric model. As such, we conjecture that our framework can be adapted to those settings as well. We leave that to future research.

literature and to additive models, where $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = G(\theta_{01}(\boldsymbol{x}_1) + \theta_{02}(\boldsymbol{x}_2))$, for non-overlapping subsets $\boldsymbol{x}_1, \boldsymbol{x}_2$ of $\boldsymbol{x}$, we will obtain rates for $\theta_{01}(\boldsymbol{x}_1)$ and $\theta_{02}(\boldsymbol{x}_2)$.

The form of Equation (2.3) makes clear that our approach is to *enrich* a parametric relationship between the outcomes $\boldsymbol{Y}$ and policy variables $\boldsymbol{T}$, rather than *restrict* a fully nonparametric (prediction) model. It is better to view $G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t})$ as an ML-enriched version of $\mathbb{E}[Y \mid \boldsymbol{T} = \boldsymbol{t}] = G(\boldsymbol{\theta}_0'\boldsymbol{t})$ instead of as a restricted version of the a prediction model such as $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = \theta_0(\boldsymbol{x}, \boldsymbol{t})$, which would be more typical of ML.

This distinction is both practically and theoretically important. Again, consider the binary choice model of our application. In our framework, the heterogeneous interest rate (price) effect is directly available as a coefficient function. Compare this to the unstructured prediction case: to recover the same conceptual quantity one would first estimate the prediction function $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = \theta_0(\boldsymbol{x}, \boldsymbol{t})$ and then obtain the derivative with respect to the rate variable R (an element of $\boldsymbol{T}$), that is, $\mathbb{E}[\partial\theta_0(\boldsymbol{x}, \boldsymbol{T})/\partial R|\boldsymbol{X} = \boldsymbol{x}]$. This is possible, but cumbersome, and inference on the average may not be regular without weighting (see also Section 6.3). It would be difficult or impossible to recover measures such as elasticities or optimal prices. However, with our structural model, all of these are simple and automatic.

3 Structured Deep Learning for Parameter Functions

We now discuss in detail the deep neural network (DNN) estimation of the parameter functions $\boldsymbol{\theta}_0(\boldsymbol{x})$ and state our theoretical results for the first stage. DNNs have a unique combination of strengths which makes them an excellent choice, among machine learning methods, for recovering individual heterogeneity. The most obvious argument for deep learning is the incredible success DNNs have had across a wide variety of learning problems, in research applications and real-world use. They have been found to handle many different tasks and data types extremely well. In many applications the dimension of $\boldsymbol{X}$ is large enough, and the heterogeneity complex enough, that classical methods will not work, but deep learning still yields excellent results. See Goodfellow et al. (2016) for a textbook treatment and Farrell et al. (2021) for recent literature and further introduction.

The primary use of DNNs has been for prediction, and much of the statistical study has been restricted to this case. We take DNNs beyond prediction, and use them to learn the (structural)

parameter functions $\boldsymbol{\theta}_0(\boldsymbol{x})$. To do so, we design a new architecture, shown in Figure 1, to measure the loss directly in terms of the parameter functions. The key idea is to decouple the final loss and the functions to be learned: we use DNNs to approximate the parameter functions $\boldsymbol{\theta}_0(\boldsymbol{x})$ in a penultimate "parameter layer", and these are then combined according to the economic model in the final, "model layer" of the network. This is crucial because it forces the machine learning to be faithful to the economic structure and it allows us to learn the components of $\boldsymbol{\theta}_0(\boldsymbol{x})$, which are of direct interest and required for learning $\boldsymbol{\mu}$. To approximate the functions we use standard fully-connected feedforward networks (multi-layer perceptrons, MLPs) and the rectified linear unit (ReLU) activation function.

This change, from θ to $\theta(\boldsymbol{X})$ is simple, yet powerful. It allows for deep learning of individual heterogeneity on any interesting parameter function. These functions may be coefficients that ultimately go into a prediction, as in Equation (2.3), variances or covariances, or other parameters of the original model. To make this formal, let $\mathcal{F}_{\text{DNN}}$ be a class of ReLU-DNNs that is restricted to our structured architecture, and also yields $\|\theta_k\|_\infty$ bounded. We then obtain $\widehat{\boldsymbol{\theta}}$ by solving the empirical analogue of (2.1),

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathcal{F}_{\text{DNN}}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \ell\left(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{\theta}(\boldsymbol{x}_i)\right). \tag{3.1}$$

DNNs are not the only possible method that can be used, nor do we claim any formal optimality for them. Having said that, from a practical point of view, they are ideal for several reasons. First, we are able to easily and transparently make our estimator mirror the global structure of the model because DNNs learn a basis-function style representation, and in this way are akin to a global smoother.[3] More apt for the present purpose, we dub this property "structural compatibility". This makes the machine learning faithful to the economics, rather than allowing the reverse. Although it is possible to embed "local" methods, such as kernel-based (Fan and Zhang, 2008) or tree-based (Zeileis et al., 2008; Athey et al., 2019; Chatla and Shmueli, 2020) estimators, doing so with DNNs is simple, transparent, and tractable, and as the economic model holds globally, we may wish to match this in estimation. Second, like tree-based methods, DNNs handle discrete covariates automatically,

[3]The distinction between a global and local smoother is not universal or precise. Here, we use "global" to mean that the estimator imposes a global smoothing across the data, typified by a series estimator, whereas a "local" estimator imposes only local smoothing structure, typified by Nadaraya-Watson kernel regression. However, this need not match the notion of using global versus local data in estimation: for example, although series estimators are generally regarded as global, those such as splines or partitioning, through their specific basis functions, use only local data (Cattaneo and Farrell, 2013; Cattaneo et al., 2020b).

including fully flexible interactions. We do not need to restrict attention, in practice or in theory, to continuous $\boldsymbol{X}$. In nonparametric theory there is often no penalty in the rate of convergence for discrete variables (under standard assumptions), but realizing these gains in practice can be difficult, as most nonparametric estimators are designed with continuous variables in mind (such as those built upon basis expansions or kernel approximations). DNNs require no customization: the inputs shown in Figure 1 can be any mix of continuous and discrete data. By exploiting our structured architecture, we prove that only the dimension of the continuous elements of $\boldsymbol{X}$, the heterogeneity, affects the rates of convergence of our DNN estimators, neither discrete covariates nor $d_{\boldsymbol{T}}$ impact the rate.

The handling of discrete covariates is a major advantage over classical sieve methods or methods that select series terms, such as the lasso. Classical methods are ill-equipped to handle modern complex, high-dimensional tasks, and, in practice, variable selection methods require pre-specifying the functional forms and interactions over which to search. However, such methods are structurally compatible in our sense, and our results contribute directly to the large body of work on nonparametric M-estimation (see Chen (2007) discussion and further references). Setting our methodology and theory apart from earlier work is that we explicitly advocate the enrichment of a standard structural model, relating $\boldsymbol{Y}$ and $\boldsymbol{T}$, rather than starting with a fully generic case. The more common starting point in semi- or nonparametric M estimation and inference would be a model such as $\ell\left(\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\theta}(\cdot)\right)$, instead of our explicit enriched model and two step approach. Our approach is what allows us to deliver concrete, fully implementable results. Extensions to more general settings would be a useful future step, however.

Finally, it is worth mentioning that other recent work has considered the combination of deep learning and some form of structural modeling (examples include Wei and Jiang, 2019; Igami, 2020; Kaji et al., 2020; Chen et al., 2021). Typically, the goal is estimation of a parametric structural model and deep learning methods are applied to learn the mapping of data to parameters. Our focus, using deep learning to estimate individual-level heterogeneity, is quite different, and further, we give theoretical results on deep neural network estimation and subsequent inference which are not available in prior work. Babii et al. (2020) combine economics with prediction-focused machine learning, where economic reasoning is used to support assymetric in classification loss functions. This is another interesting avenue for merging economics with machine learning.

3.1 Convergence for Structured Deep Neural Networks

We now state our theoretical results for DNN estimation of the parameter functions $\boldsymbol{\theta}_0(\boldsymbol{x})$. We will give a generic result, which necessarily requires high level conditions, and then illustrate more concrete ideas in the case of the slope and intercept parameter functions. The results in this section contribute to the recent literature on the statistical properties of deep learning and to the longer tradition of sieve M estimation. Our assumptions and results are reminiscent of both.

We require two sets of conditions: one covers the model $\ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ and one gives regularity for $(\boldsymbol{Y}, \boldsymbol{X}', \boldsymbol{T}')'$ and $\boldsymbol{\theta}_0(\boldsymbol{x})$. 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 (2.1), uniformly bounded, and that there are constants c_1, c_2, and C_ℓ that are bounded and bounded away from zero, such that*

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

The curvature requirement will often be implied by restrictions on the Hessian or on the matrix $\boldsymbol{\Lambda}(\boldsymbol{x})$. Such restrictions are natural in our setting, as they will be required for inference anyway, and are known to hold in many contexts. Some potentially interesting cases are ruled out, such as quantile regression. Modifications to our theory could allow for these cases: for example, Padilla et al. (2020) apply the methods of Farrell et al. (2021) to quantile regression.

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.*

We now have the following result, proven in the appendix. Here we focus on smooth functions as well as the commonplace deep and wide multi-layer perceptrons (fully connected, feedforward neural networks). In the appendix, we give a more general result, one that is agnostic about the type of approximation, and hence the type of network. For example, the general results can be used to obtain faster rates or cover fixed-width, very deep networks (Farrell et al., 2021, Section 2.3).

Theorem 1. *Let* $\boldsymbol{w}_i$, $i = 1,\ldots,n$, *be a random sample that obeys Assumptions 1 and 2. For* $\widehat{\boldsymbol{\theta}}$ *solving (3.1), with* $\mathcal{F}_{\text{DNN}}$ *structured according to Figure 1, with width* $H \asymp n^{(d_C)/2(p+d_C)} \log^2 n$ *and depth* $L \asymp \log n$, *it holds that*

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

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 dimensions* $\boldsymbol{W}$, $d_{\boldsymbol{\theta}}$, *and other fixed quantities in Assumptions 1 and 2.*

The result of Theorem 1 speaks directly to the nonparametric M estimation literature. It shows that deep nets enjoy the same properties of other methods, but has the advantages discussed above. A theoretical drawback is that for a given smoothness level, this rate is not optimal. It is however sufficiently fast for inference and reflects the excellent empirical performance. This theorem fully takes deep learning away from prediction and toward learning economically meaningful parameters.

The conditions of Theorem 1 are necessarily high level, giving the generality of the setting. Verying these conditions in a given setting is required, but is often straightfoward. Importantly,

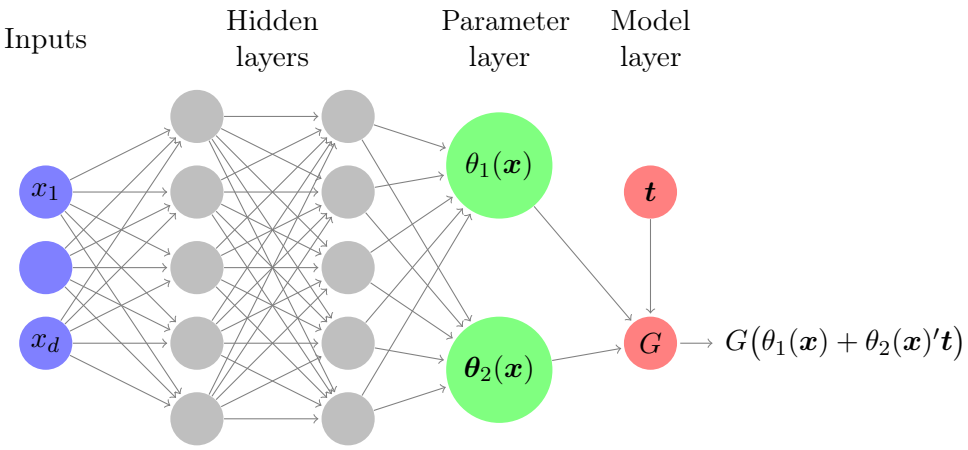


Figure 2: Illustration of the deep neural network estimation of the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ for a generic structured model (2.3)

because we are specifically focusing on enriching a parametric 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.

To illustrate, return to the regression-type model of (2.3), where the model is $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t})$ and $\boldsymbol{t}$ includes a constant term. The final loss in this case still revolves around prediction (beit a QMLE or nonlinear least squares), but our architectural idea is still important, because we need to learn the slope and intercept functions separately. This is shown in Figure 2. That figure also illustrates how our results apply immediately 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.

For this model we can also illustrate the verification of the high level conditions with familiar, primitive assumptions in this case.

Assumption 3. *(i) The conditional expectation $G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t})$ 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. For example, consider condition (iii) in the original model $\mathbb{E}[Y \mid \boldsymbol{T} = \boldsymbol{t}] = G(\boldsymbol{\theta}_0' \boldsymbol{t})$. In such a model, the assumption that $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}']$ is positive definite would be standard. Leaning on the intuition that the enriched model is akin to running the original model for each $\boldsymbol{X} = \boldsymbol{x}$, we see the condition is exactly what would be expected. Versions of these assumptions are also required in the semiparametric models that are special cases of $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G(\boldsymbol{\theta}_0(\boldsymbol{x})' \boldsymbol{t})$, including treatment effects and partially linear models, as in Section 6.

These assumptions are sufficient for identification of the slope and intercept functions in this case. A similar verification should be done for other models, and may lean on prior work in parametric M estimation, as discussed for several cases in Section 6. Specializing Theorem 1 to this case, we have the following result.

Corollary 1. *Let the conditions of Theorem 1 and Assumption 3 hold. Then for a DNN structured according to Figure 2,* $\|\widehat{\theta}_k - \theta_{0k}\|^2_{L_2(\boldsymbol{X})} = O(n^{-\frac{p}{p+d_C}} \log^8(n))$*, for* $k = 1, \ldots, d_{\boldsymbol{\theta}}$*, and* $\|G(\widehat{\boldsymbol{\theta}}(\boldsymbol{x})' \boldsymbol{t}) - G(\boldsymbol{\theta}_0(\boldsymbol{x})' \boldsymbol{t})\|^2_{L_2(\boldsymbol{X})} = O(n^{-\frac{p}{p+d_C}} \log^8(n))$.

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 $\mathbb{E}[Y|\boldsymbol{x}, \boldsymbol{t}]$ as a whole.

We also state a result for estimating the regression function $\mathbb{E}[Y|\boldsymbol{x}, \boldsymbol{t}]$ in the structural model (2.3). From a statistical point of view, this result establishes that structured DNNs have excellent performance in varying coefficient models, additive models, and other such cases, and therefore may be of independent interest. 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. The key feature to note is that the convergence rate depends only on the dimension of (the continuous component of) $\boldsymbol{X}$, not $d_{\boldsymbol{T}}$. Even if the final inference relies on conditional expectations, or if the parameter functions could be obtained from the unstructured predictions, naive estimation of these would give a much slower rate, dependent on $d_{\boldsymbol{X}} + d_{\boldsymbol{T}} - 1$ (since $\boldsymbol{T}$ includes an intercept). This would often be too slow for subsequent inference. For example, in our empirical illustration this would require 22 dimensional nonparametrics, which may be prohibitively high even for deep learning, and if the goal is to recover

a measure of "linear" impact, such as a treatment effect, marginal effect, or other average derivative, an unnecessary complication.

That the rate depends only on $d_{\boldsymbol{X}}$ is intuitively exactly what should happen: the heterogeneity is where the model is flexible. In our case, this result is due to our structured architecture, Figures 1 and 2, and *not* adaptive estimation. The model structure is enforced, not recovered. Certain specialized types of DNNs may in fact adapt to such structures (Bach, 2017; Bauer and Kohler, 2019; Schmidt-Hieber, 2019). First, this is not useful for our purposes because $\boldsymbol{\theta}_0(\boldsymbol{x})$ cannot be recovered. Second, experience has shown that imposing the structure of the model (2.1) improves estimation quality when the structure exists to allow for adaptivity.

For other recent theoretical results on deep learning in other contexts, see Liang (2018), Polson and Ročková (2018), Wang and Ročková (2020), Liang and Tran-Bach (2020), and references therein as well as in Farrell et al. (2021). One important aspect we do not address is regularization, neither the implicit regularization that may occur in the optimization nor explicit regularization of the network parameters themselves though norm penalties, weight decay, drop out, or other methods. In our applications we obtain excellent performance without using explicit regularization, though in low signal-to-noise scenarios adding explicit regularization to the implicit regularization may yield improvements, and has been shown to be optimal in an adversarial game with nature (Blanchet et al., 2020). The role of regularization, its implementation, and its consequences for estimation and subsequent inference, are major open questions for deep learning.

4 Influence Function and Semiparametric Inference

With the framework in place and estimates of the parameter functions in hand, we now turn to estimation and inference for $\boldsymbol{\mu}_0 = \mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)]$ from (2.2). The key contribution here, given in Theorem 2, is a novel Neyman orthogonal score for any such $\boldsymbol{\mu}_0$, given any model (2.1). Importantly, this score can be immediately taken to data, without additional derivations. Our framework is simultaneously general enough to cover a large number of settings, yet specific enough that the influence function can be fully characterized. As is standard, the orthogonal score depends on $\boldsymbol{\theta}_0(\boldsymbol{x})$ and a second nonparametric object, but the latter piece is also fully characterized and is therefore estimable without additional work. With this score, we can apply the methods and results

of Chernozhukov et al. (2018) to immediately obtain asymptotic Normality. This is spelled out in Section 4.2 below.

4.1 Influence Function

Obtaining valid semiparametric inference does not require basing the estimation framework on an influence function (or Neyman orthogonal, doubly robust, or locally robust, scores). The major appeal of these methods is that we can obtain valid distributional approximations under weaker conditions on the first stage estimates, i.e., on how well $\widehat{\boldsymbol{\theta}}$ recovers $\boldsymbol{\theta}_0$. These weaker conditions are known to hold for many ML methods, and in particular Section 3 shows that they hold for deep learning. In other words, it is useful to view the influence function as a tool for obtaining feasible inference, rather than being of direct interest itself, for efficiency considerations or other comparisons. This viewpoint is implicit in much recent work on inference after ML (e.g., Belloni et al., 2014; Farrell, 2015; Chernozhukov et al., 2018) but it is worthwhile to make it explicit to better understand how thinking this way greatly expands the breadth of what we can cover, including cases like the optimal interest rate, which cannot be given in closed form. The same motivation is explicit in the recent work on "auto-DML", where the necessary pieces of the influence function are estimated from the data, and are thus need not be derived at all (Chernozhukov et al., 2020c,b,a,d, 2021). Our focus on interpretable parameter functions rather than regression functions distinguishes our setting from this line of work, which manifests in two central ways: (i) our first stage is more general, as we do not focus on regressions; (ii) our inference targets are broader in some ways, given the flexibility of the first stage, but we require $\boldsymbol{\theta}_0$ to enter through evaluation at $\boldsymbol{X}$, ruling out examples such as integrals across data points.

Influence functions have a long history in econometrics. Newey (1994) remains the seminal treatment. We defer to that work and Ichimura and Newey (2015) for more background, including regularity conditions for existence of an influence function. Perhaps the most well known and commonly used influence function is that for average treatment effects, which we recover as a special case in Example 6.1. 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 certain plug-in estimators could be efficient (Hirano et al., 2003; Imbens et al., 2007), and finally only recently used for post-ML inference to exploit the weaker conditions (Farrell, 2015). The partially

linear model (Section 6.3) is another standard example and followed a similar trajectory, most imporantly for our discussion being the first setting where inference was proven (uniformly) valid after variable selection (Belloni et al., 2014).

In both cases, the influence function was first derived and then its components estimated directly. This exercise has been repeated in numerous models for numerous parameters (see Section 6, and lists in Ichimura and Newey (2015) or Chernozhukov et al. (2020a)). In a sense, we follow this path: derive the correction factor for any setting in Section 2, obeying the regularity conditions below. However, our correction factor covers many settings, as we take advantage of the structure of the original model in the derivation. Thus, in any of these settings, inference is now as straightforward as it is for average treatment effects: two nonparametric pieces must be estimated, and these are combined into the orthogonal score, with sample splitting if needed. That is, we can follow the recipe of Chernozhukov et al. (2018). We broaden the application of these ideas and show how they work even in cases where we use automatic differentiation to obtain the piece of the score and the parameter of interest is not available in closed form. That is, we can still evaluate the influence function estimator, given below, on the data points. And so, while deriving correction terms for estimation is somewhat standard, or results are novel in their broad applicability. Our influence function nests many cases from the literature and delivers new results.

Our end result will reflect the core idea behind our framework: enriching a standard model by converting the parameters to parameter functions. In purely parametric two-step models, without heterogeneity, the influence function of the first step parameters themselves can be used to adjust the second step (Newey and McFadden, 1994, Section 6). Our result is the nonparametric generalization of this, and one strength of the result is this familiarity, which will aid in practice by making assumptions transparent and familiar and by guiding implementation.

To state the result, we will in fact require the gradient and Hessian of ℓ, defined via *ordinary* differentiation. Let $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x}))$ be the $d_{\boldsymbol{\theta}}$-vector of first derivatives,

$$\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) = \left. \frac{\partial \ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b})}{\partial \boldsymbol{b}} \right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x})}, \tag{4.1}$$

and $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ as the $d_{\boldsymbol{\theta}} \times d_{\boldsymbol{\theta}}$ matrix of second derivatives, that is with $\{k_1, k_2\}$ element given by

$$\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))\right]_{k_1,k_2} = \left.\frac{\partial^2 \ell\left(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{b}\right)}{\partial b_{k_1} \partial b_{k_2}}\right|_{\boldsymbol{b}=\boldsymbol{\theta}(\boldsymbol{x})}, \tag{4.2}$$

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.

We can now state our assumptions and give the main result of this section, the form of the orthogonal score. Our assumptions are mostly standard and ensure sufficient regularity for our influence function to be calculated and asymptotic normality obtained by resulting estimator. One conceptually important point is that in general these conditions are not sufficient for a causal interpretation, which will require some form of unconfoundedness or conditional exogeneity (as in Examples 6.1 or 6.3).

Assumption 4. *The following conditions hold on the distribution of $\boldsymbol{W} = (Y, \boldsymbol{X}', \boldsymbol{T}')'$, uniformly in the given conditioning elements. (i) Equation (2.1) 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) For $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ of (4.2), $\boldsymbol{\Lambda}(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{x}))|\, \boldsymbol{X} = \boldsymbol{x}]$ is invertible with bounded inverse. (iv) The parameter $\boldsymbol{\mu}_0$ of Equation (2.2) 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 variance.*

The most important assumptions here are that the first order condition of (2.1) holds and that $\boldsymbol{\mu}_0$ is pathwise differentiable. The latter keeps focus on 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. Of course not all models of the form (2.1) will be so identified, and this must be verified.

For intuition, consider the case of the conditional mean restriction (2.3) where simple sufficient conditions can be stated. Many loss functions yield $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = \boldsymbol{t}(G(\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t}) - y)$, in which case $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\dot{G}\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X} = \boldsymbol{x}]$, with $\dot{G}$ the derivative of G with respect to its scalar argument, evaluated at the index $\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}$, $\dot{G} = (dG/du)|_{u=\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{T}}$. Condition (iii) will then often be implied by other conditions on the model. For example, if G is the logistic link and $\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}]$ 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.

We can now state our influence function result, derived in Appendix B.

Theorem 2. *Let Assumption 4 hold. Recall the definitions of* $\boldsymbol{\ell}_{\boldsymbol{\theta}}$ *in* (4.1) *and* $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))|\, \boldsymbol{X} = \boldsymbol{x}]$ *for* $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ *of* (4.2). *Define* $\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)$ *as the* $d_{\boldsymbol{\mu}} \times d_{\boldsymbol{\theta}}$ *Jacobian of* $\boldsymbol{H}$ *with respect to* $\boldsymbol{\theta}$, *that is, the matrix with* $\{h, k\}$ *element, for* $h = 1, \ldots, d_{\boldsymbol{\mu}}, k = 1, \ldots, d_{\boldsymbol{\theta}}$, *given by*

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

with H_h *the* h^{th} *element of* $\boldsymbol{H}$ *and* b_k *the* k^{th} *element of* $\boldsymbol{b}$. *Then for* $\boldsymbol{\mu}_0$ *of Equation* (2.2), *a valid and Neyman orthogonal score is* $\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_0$, *where*

$$\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*\right) - \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})). \tag{4.3}$$

At this level of generality, this influence function is new to the literature and yields many new contexts for inference after ML. In some special cases we recover existing results, particularly under (2.3) with $G(u) = u$, such as average treatment effects (Section 6.1), partially linear models (Section 6.2), and average partial effects (Section 6.3). Most importantly, when the first stage is restricted to regression under the squared loss, Newey (1994) gives the form of the correction factor for a broader set of moment conditions than (2.2). Beyond these specific examples, our result appears new, both in generality and in the many concrete cases we give, such as choice models or linear IV models.

The form of the influence function is standard: a plug-in piece and a correction or debiasing piece. The correction term relies on three derivatives, $\boldsymbol{H}_{\boldsymbol{\theta}}$, $\boldsymbol{\ell}_{\boldsymbol{\theta}}$, and $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}$, and these can be computed easily using automatic differentiation, if they are unknown, and thus there is no derivation required before estimation can take place.

The form of the correction term, specifically $\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$, warrants further discussion. Behind this form is again the fact that we start with a model and enrich its parameters to be functions. This means that the parametric submodels that lie behind the pathwise derivative calculation simply trace through the space of the original, parametric structural model, which is well understood and

well behaved. The fact that the model relates $\boldsymbol{Y}$ and $\boldsymbol{T}$ in a known way, with parameters enriched using $\boldsymbol{X}$, is important as it allows for a simple isolation of the contribution of the nonparametric estimation. Due to this, and the two-step nature of our set up, this term does not depend on $\boldsymbol{H}$, which is helpful when several parameters are of interest in one application.

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. The presence of an inverse function is commonplace in semiparametric inference problems. The $\boldsymbol{\Lambda}(\boldsymbol{x})$ matrix is never high dimensional, again due to our approach (cf Remark 1). Estimation or calculation of $\boldsymbol{\Lambda}(\boldsymbol{x})$ will simplify in many cases, particularly if one has prior knowledge that only a certain subset of the heterogeneity covariates, say $\boldsymbol{X}_1 \subset \boldsymbol{X}$, are relevant for $\boldsymbol{T}$. An extreme case is randomization (Remark 4 below). Another example occurs with $\boldsymbol{T}$ being prices, which are known with certainty to be set by the firm according to only several characteristics of the market or consumer.

One appealing aspect of our influence function is that we do *not* have a nonparametric (conditional) density function. The matrix $\boldsymbol{\Lambda}(\boldsymbol{x})$ consists only of regression-type objects: we must project derivatives of the loss onto $\boldsymbol{X}$. Again we can use treatment effects for intuition: we know exactly what nonparametric regression object is required, the propensity score, and we must estimate it to form the empirical influence function in observational data or we can compute it in experiments. Our result is at the same level, given estimates $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$, though it may appear more cumbersome in terms of actual coding. We discuss estimation in more detail in Section 4.2 and implementation in Section 5.

Two other parallels with treatment effects are worth noting regarding $\boldsymbol{\Lambda}(\boldsymbol{x})$. First, in many problems, if $\boldsymbol{T}$ is randomized, estimation is not required and $\boldsymbol{\Lambda}(\boldsymbol{x})$ can be computed directly (Remark 4), just as the propensity score need not be nonparametrically estimated in experiments. Second, as is standard in semiparametrics, ensuring that the "denominator" is bounded away from zero is crucial, which is the impetus behind the prevalent trimming on the propensity score. Although our results allow for estimation using the influence function without knowing its precise form, this is not always desirable in practice, as we may wish to trim based on more primitive objects if possible. In treatment effects, we cannot trim based on the propensity score unless that piece of the influence function is known. Trimming based on the propensity score is cleanly interpretable based on the overlap condition (Crump et al., 2008) and can be studied theoretically (Ma and Wang, 2020),

and is therefore more appealing that regularization that is not grounded in economics. In practice, trimming or other regularization of $\boldsymbol{\Lambda}$, such as using $(\boldsymbol{\Lambda}(\boldsymbol{x}) + \boldsymbol{I}_{d_{\boldsymbol{\theta}}})^{-1}$, may be helpful.

Remark 1. Our derivation deals directly with the nonparametric objects $\boldsymbol{\theta}_0(\boldsymbol{x})$ and we obtain a result familiar from parametric two-step models (Newey and McFadden, 1994). This can be contrasted with a seemingly-similar approach to inference that treats also evokes parametric two-step models. Here, the first stage parameters are those of the nonparametric estimator itself. If $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ were a series estimator, for example, the parameters would be the coefficients on the basis functions. In the case of lasso, the parameters match the high-dimensional variables. One can then treat this as a (large) two-step parametric problem to obtain valid inference, as in Ackerberg et al. (2012). Applying this idea to deep learning is in some ways tempting, because fitting DNNs is maximum likelihood, which is in principle well understood. However, pursuing this approach would lead to impractical results due to the high dimensionality of modern ML methods. For example, the equivalent of $\boldsymbol{\Lambda}(\boldsymbol{x})$ in this case 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 would be challenging or impossible practically and potentially invalid theoretically, and moreover, given our results, it is unnecessary. ♣

Remark 2. The goal of Theorem 2 is to allow for *feasible* inference post-ML in a wide variety of empirically useful settings, rather than explicitly targeting *efficient* inference. However, in many cases our result matches the efficient influence function, such as for average treatment effects or heterogeneity-enriched linear models more generally. Interestingly, although (2.3) is more flexible than the usual partially linear model, which has constant slopes, we recover the usual result of efficiency under homoskedasticity in linear models, but not in nonlinear models. All of these are discussed in Section 6. We conjecture that our result yields efficiency more broadly, when the original model is based on a likelihood or exponential family, by arguing as in Remark 4.1 of Mammen and van de Geer (1997). ♣

Remark 3 (Orthogonal Loss Functions)**.** We use an orthogonal score for inference only, where the orthogonality is needed to ensure validity after learning the parameter functions $\boldsymbol{\theta}_0$ in the first

stage. An alternative approach recently considered is to exploit orthogonality in the estimation stage as well, or to perform these jointly. Foster and Syrgkanis (2020) used orthogonal scores to study M estimation problems with a nuisance parameter, i.e., $\ell(\boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\theta})$ in our notation. Tan (2020) used a modified loss function to obtain doubly robust inference on average treatment effects with high-dimensional sparse models, compared to the "consistency" type of robustness obtained by using the influence function based estimators. Nekipelov et al. (2020) considered a model like (2.3), but with high-dimensional sparse linear models, so that $\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t} = (\boldsymbol{\theta}_0'\boldsymbol{x})'\boldsymbol{t}$ where many entries of vector $\boldsymbol{\theta}_0$ are zero. Under the motivation that the "complexity of the control function is likely to be much larger than [that] of heterogeneous interactions", they develop a loss function for estimation of the vector $\boldsymbol{\theta}_0$ that is automatically orthogonal to estimation of other nuisance parameters. Similarly, in the context of heterogeneous effects of a binary treatment, where (2.3), under a linear link and binary scalar $\boldsymbol{t}$, is without loss of generality, Nie and Wager (2020) and Kennedy (2020) (and references therein) develop estimation procedures for $\theta_{02}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = 1] - \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = 0]$ which obtain better rates if the function $\theta_{02}(\boldsymbol{x})$ has a simpler structure than $\theta_{01}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = 0]$. Common to these approaches is that the heterogeneity is in some way "simpler" than the rest of the problem, which is the opposite of our goal of capturing rich individual heterogeneity. ♣

It will be useful to explicitly state how Theorem 2 applies to the regression models (2.3). We will use this form in several examples in Section 6 and in two remarks below.

Corollary 2. *Assume the model* (2.3) *and that the loss is such that* $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = \boldsymbol{t}(G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}) - y)$ *and therefore* $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\dot{G}\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X} = \boldsymbol{x}]$*, where* $\dot{G} = (dG/du)|_{u=\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{T}}$*. Then under the conditions of Theorem 2, the result there holds with*

$$\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = \boldsymbol{H}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*) + \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)\mathbb{E}[\dot{G}\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X} = \boldsymbol{x}]^{-1}\boldsymbol{t}(y - G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t})). \tag{4.4}$$

Here we can see exactly that $\boldsymbol{\Lambda}(\boldsymbol{x})$ requires projecting weighted first and second moments of $\boldsymbol{T}$. This is intuitive from linear models, where the conditional variance is the crucial object (Sections 6.1, 6.2, 6.3) and from other applications of generalized linear models and QMLE.

The following two remarks use Corollary 2 for two cases where $\boldsymbol{\psi}$ simplifies, which are common enough in applications to be worth spelling out here.

Remark 4 (Randomized Treatments)**.** If $\boldsymbol{T}$ is randomly assigned, or more generally is independent of $\boldsymbol{X}$, then $\boldsymbol{\Lambda}(\boldsymbol{x})$ often simplifies or can be directly computed. In many problems of interest, $\boldsymbol{\ell_{\theta\theta}}$ will not be a function of $\boldsymbol{y}$, only $\boldsymbol{t}$ and (through $\boldsymbol{\theta}_0$) $\boldsymbol{x}$. Thus under randomization, given $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}_i)$, $\boldsymbol{\Lambda}(\boldsymbol{x}_i)$ can be computed and need not be estimated, though it remains a function of $\boldsymbol{x}$ in general, as opposed to simply a constant. In the case of (4.4), $\boldsymbol{\Lambda}(\boldsymbol{x}) = \int \dot{G}(\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t})\boldsymbol{t}\boldsymbol{t}'dF_{\boldsymbol{T}}(\boldsymbol{t})$. If the distribution, denoted $F_{\boldsymbol{T}}(\boldsymbol{t})$, is known, this object can be computed numerically for fixed functions $\boldsymbol{\theta}(\boldsymbol{x})$ (or if $G(u) = u$, these are not needed). This motivates the three-way sample splitting for nonlinear models discussed below. Note that this continues to apply under cases such as stratified randomization, where the relevant distribution will be known and depend on a very simple subset of the covariates. ♣

Remark 5 (Scalar Parameters with Univariate Treatments)**.** To state a simple, concrete result, consider the case where $\boldsymbol{\mu}_0$ is scalar and (2.3) holds with scalar treatment variable, so that $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, T = t] = G(\theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x}) \cdot t)$. Then we can invert $\boldsymbol{\Lambda}(\boldsymbol{x})$ manually. The (scalar) function $\psi(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda})$ in this case is more familiar and ease to compare to earlier results. Define $\lambda_k(\boldsymbol{x}) = \mathbb{E}[\dot{G}T^k|\boldsymbol{X} = \boldsymbol{x}], k = \{0, 1, 2\}$, $\dot{H}_1(\boldsymbol{x}) = \partial H(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)/\partial\theta_{01}$, and $\dot{H}_2(\boldsymbol{x}) = \partial H(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^*)/\partial\theta_{02}$. Then Theorem 2 holds with

$$\begin{aligned}\psi(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = {} & H\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); t^*\right) \\ & + \frac{\dot{H}_1(\boldsymbol{x})\left(\lambda_2(\boldsymbol{x}) - \lambda_1(\boldsymbol{x})t\right) + \dot{H}_2(\boldsymbol{x})\left(\lambda_0(\boldsymbol{x})t - \lambda_1(\boldsymbol{x})\right)}{\lambda_2(\boldsymbol{x})\lambda_0(\boldsymbol{x}) - \lambda_1(\boldsymbol{x})^2}\Big(y - G(\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t})\Big).\end{aligned} \tag{4.5}$$

We note here that the treatment may be discrete or continuous, in either case the form of the influence function remains the same. Again this result can be used to compare to other cases and ease implementation. ♣

4.2 Asymptotic Normality

With the orthogonal score of Theorem 2 in hand, we now turn to point estimation and inference for $\boldsymbol{\mu}_0$. We will apply the methods and results from Chernozhukov et al. (2018) and as such we keep the discussion brief. The crucial point here is that in order to form the point estimator $\widehat{\boldsymbol{\mu}}$ and standard errors $\widehat{\boldsymbol{\Psi}}$ we need only to evaluate the influence function at each data point, and this can be done

in the full generality of Theorem 2. This is still possible when, as mentioned above, the required elements are not available in closed form.

We will rely on sample splitting or cross fitting here, in order to obtain the desired theoretical result. From a theoretical point of view, sample splitting allows us to obtain a properly centered limiting distribution under weaker conditions on the first stage (deep neural network) estimates for all applications of our framework. However, from a practical point of view, sample splitting or cross fitting come with a cost that can be large in some applications. One obvious cost is computational: the machine learning must be done multiple times on the different subsamples. A more subtle cost, but crucial when sample sizes are small, is that the smaller (sub-)sample sizes can yield worse results. Sample splitting relies on the asymptotic fact for fixed S, a sample of size n/S is as good as the full sample of size n. In practice, this may not hold, particularly for challenging nonparametric estimation problems. Farrell et al. (2021) show that for some estimands sample splitting is not needed for inference after deep learning under standard assumptions. It would be useful to extend that argument to more general settings.

For estimation of $\boldsymbol{\Lambda}(\boldsymbol{x})$ we may in fact require three-way splitting. Because the "outcome" required for these projections depends on the unknown $\boldsymbol{\theta}_0(\boldsymbol{x})$, we will estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ on one subsample, use these to obtain $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{x})$ on a second sample, and then use the final portion for the parametric estimation and inference. In typical cross fitting the first and second portions would be one sample. In two cases this three-way splitting is not needed. The first is under randomization when $\boldsymbol{\Lambda}(\boldsymbol{x})$ can be computed, given $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$. Second, under (2.3) when $G(u)$ is linear, $\dot{G} \equiv 1$ and $\boldsymbol{\Lambda}(\boldsymbol{x})$ is simply the covariance matrix of $\boldsymbol{T}$, conditional on $\boldsymbol{X}$, and this can be estimated along with $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$.

We will be brief in describing the estimation procedure, leaving further discussion to Chernozhukov et al. (2018) and Newey and Robins (2018). First, the observations $\{1, \ldots, n\}$ are split into S subsets, denoted by $\mathcal{S}_s \subset \{1, \ldots, n\}$, $s = 1, \ldots, S$. Let $\mathcal{S}_s^c$ be the complement of $\mathcal{S}_s$. We then, for each $s = 1, \ldots, S$, use $\mathcal{S}_s^c$ to obtain estimates of $\boldsymbol{\theta}_0(\cdot)$ and $\boldsymbol{\Lambda}(\cdot)$; denote these by $\widehat{\boldsymbol{\theta}}_s(\cdot)$ and $\widehat{\boldsymbol{\Lambda}}_s(\cdot)$. If needed, $\mathcal{S}_s^c$ is further split in two pieces, using the first to get $\widehat{\boldsymbol{\theta}}_s(\cdot)$ and the second for $\widehat{\boldsymbol{\Lambda}}_s(\cdot)$. The final estimator of $\boldsymbol{\mu}_0$ is then

$$\widehat{\boldsymbol{\mu}} = \frac{1}{S}\sum_{s=1}^{S} \widehat{\boldsymbol{\mu}}_s, \qquad \widehat{\boldsymbol{\mu}}_s = \frac{1}{|\mathcal{S}_s|}\sum_{i \in \mathcal{S}_s} \boldsymbol{\psi}(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}_s(\boldsymbol{x}_i), \widehat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}_i)), \tag{4.6}$$

where $|\mathcal{S}_s|$ is the cardinality of $\mathcal{S}_s$ and is assumed to be proportional to the sample size n. Along with the point estimator $\widehat{\boldsymbol{\mu}}$ we will require an estimator of the asymptotic variance, which is given by $\boldsymbol{\Psi} = \mathbb{V}[\boldsymbol{\psi}(\boldsymbol{W}, \boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{\Lambda}(\boldsymbol{X}))]$. To estimate $\boldsymbol{\Psi}$ we use the variance analogue of (4.6):

$$\widehat{\boldsymbol{\Psi}} = \frac{1}{S}\sum_{s=1}^{S} \widehat{\boldsymbol{\Psi}}_s, \qquad \widehat{\boldsymbol{\Psi}}_s = \frac{1}{|\mathcal{S}_s|}\sum_{i\in\mathcal{S}_s} \left(\boldsymbol{\psi}(\boldsymbol{w}_i, \widehat{\boldsymbol{\theta}}_s(\boldsymbol{x}_i), \widehat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}_i)) - \widehat{\boldsymbol{\mu}}\right)^2. \tag{4.7}$$

Asymptotic normality of $\widehat{\boldsymbol{\mu}}$ and consistency of $\widehat{\boldsymbol{\Psi}}$ will follow from Chernozhukov et al. (2018). To emphasize that our inference results, the orthogonal score in particular, are useful in semiparametrics broadly, including after ML estimation of any kind, we employ the following high-level conditions on the convergence rates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Lambda}}$.

Assumption 5. *Based on a sample of size n, the estimators for $\boldsymbol{\theta}_0$ and $\boldsymbol{\Lambda}$ obey $\|\widehat{\theta}_{k_1} - \theta_{0k_1}\|_{L_2(\boldsymbol{X})} = o(n^{-1/4})$ and $\|[\widehat{\boldsymbol{\Lambda}}]_{k_1,k_2} - [\boldsymbol{\Lambda}]_{k_1,k_2}\|_{L_2(\boldsymbol{X})} = o(n^{-1/4})$ for all $k_1, k_2 \in \{1, \ldots, d_{\boldsymbol{\theta}}\}$.*

Many nonparametric and ML estimators may satisfy this requirement. Importantly, Theorem 1 verifies these conditions for deep learning estimation of $\widehat{\boldsymbol{\theta}}$. Estimation of the elements of $\boldsymbol{\Lambda}(\boldsymbol{x})$ is a prediction problem, and therefore deep learning will also satisfy these rates, applying the results of Farrell et al. (2021) with the sample splitting discussed above. Often the squared error loss will be used here, but not always, particularly for discrete data or with nonlinear models, where a fractional outcome model or classification-based loss may be warranted.

We now have the follow result, establishing asymptotic normality and validity of standard errors. Let $\mathbf{0}_d$ be the d-long zero vector and $\boldsymbol{I}_d$ be the d-square identity matrix.

Theorem 3. *Suppose $\boldsymbol{w}_i$, $i = 1, \ldots, n$, is a random sample that obeys Assumption 4 and that Assumption 5 holds for all subsamples $s = 1, \ldots, S$, with uniformly invertible $\widehat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}_i)$. Then*

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

5 Application: Advertising and Personalized Interest Rates

5.1 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.

5.2 Model and Implementation

Our setup adapts the framework outlined in Bertrand et al. (2010) and assumes 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 distributed i.i.d. Logistic which gives the standard Logit 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 (2.1) 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 (3.1) using this $\ell(y, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ and the architecture Figure 1. 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 ($d = 13$). We use Tensorflow™ (Abadi et al., 2015) to construct the computational graph and optimize the likelihood using the ADAM optimizer (Kingma and Ba, 2014). 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 $\widehat{\boldsymbol{\mu}}$.[4]

5.3 Results and Quantities of Interest

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 explore two derived quantities. First, we 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. Second, we turn to a more ambitious goal of targeting and profit maximization, making more full use of the power of the framework.

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

[4]Complete details of the implementation are available upon request. We plan on releasing the code for the application soon.

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 and obtain inference.

5.3.1 Marginal Effects

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

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

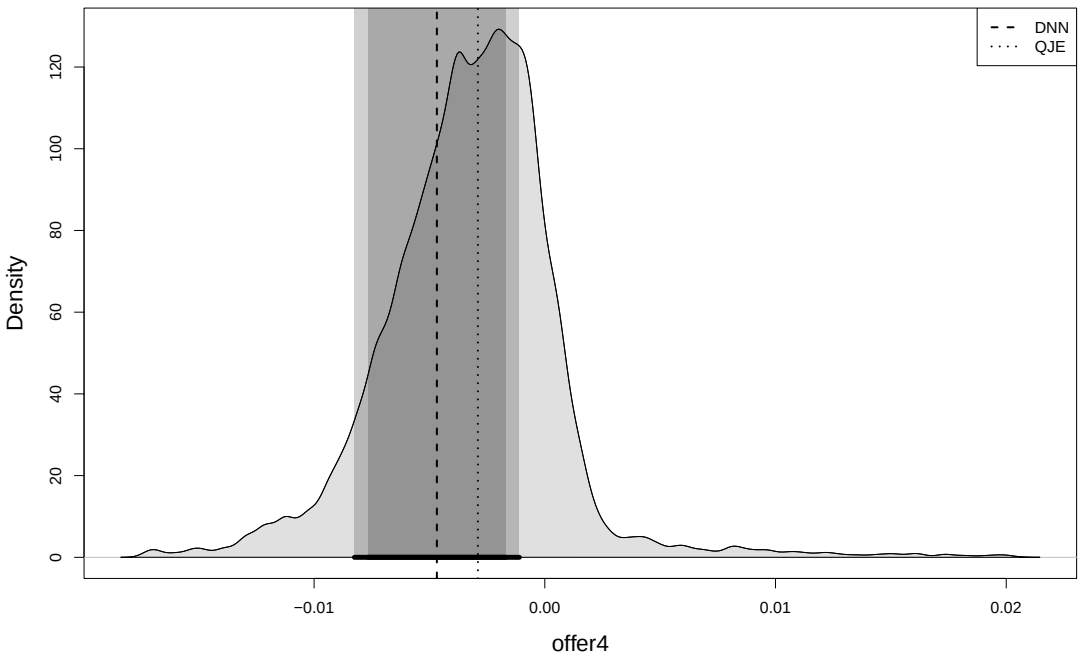
In our empirical results we set $\boldsymbol{t}^*$ to the sample average for simplicity. Similarly quantities for the advertising content are readily available. Thus, by taking $\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*) = G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}^*)(1 - G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}^*))\theta_R$ in (2.2), 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 3 and 4 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.

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

Figure 3: Marginal Effect of Interest Rate



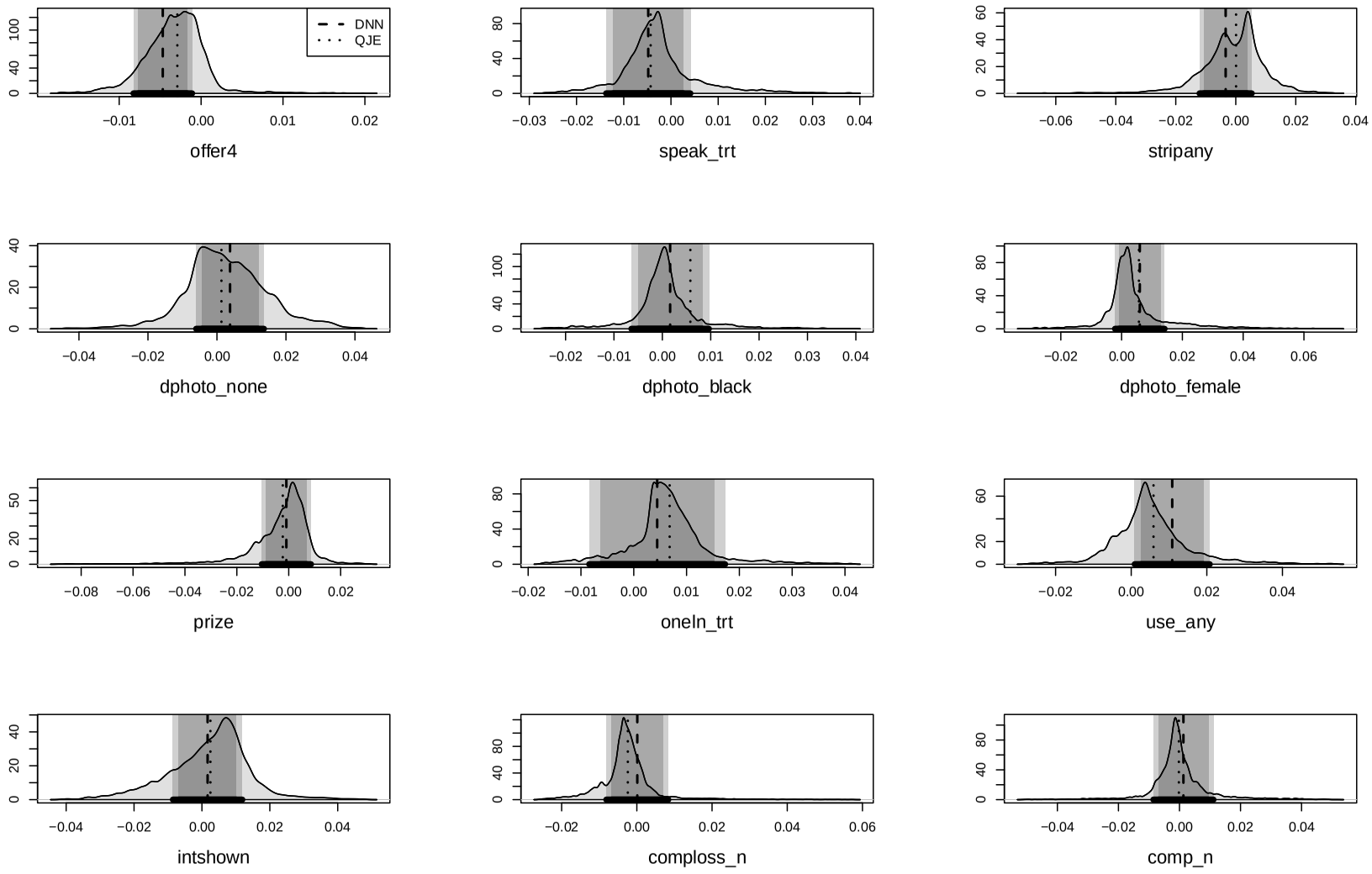


Figure 4: Marginal Effect of Advertising Content

5.3.2 Optimal Personalized Offers

Our framework allows for a rich specification of heterogeneity in the tastes of the consumer. We have shown how standard quantities of interest such as marginal effects can be easily constructed. We now demonstrate how the estimated heterogeneity can be translated into personalized offers and the simplicity with which one can conduct inference on quantities of interest (such as the mean interest rate offered or expected profits).

Given the rarity of defaults, the sample size is too small to uncover full heterogeneity, and therefore we assume that the probably of default given an interest rate $R = r$ is modeled by the function

$$\mathbb{P}[D = 1 \mid R = r] = D\left(\delta_0 + \delta_R r\right) = \frac{1}{1 + \exp\left(-\left[\delta_0 + \delta_R r\right]\right)}.$$

We take these parameters as given. To write the firm's expected profit for a given consumer, let L be the loan amount and, since we focus on optimizing the interest rate given the parameters, abbreviate $\mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G(r)$ and $\mathbb{P}[D = 1 \mid R = r] = D(r)$. Then

$$\pi(r) = L\big[rG(r)\big]\big[1 - D(r)\big]. \tag{5.2}$$

The usual optimization machinery applies and we obtain

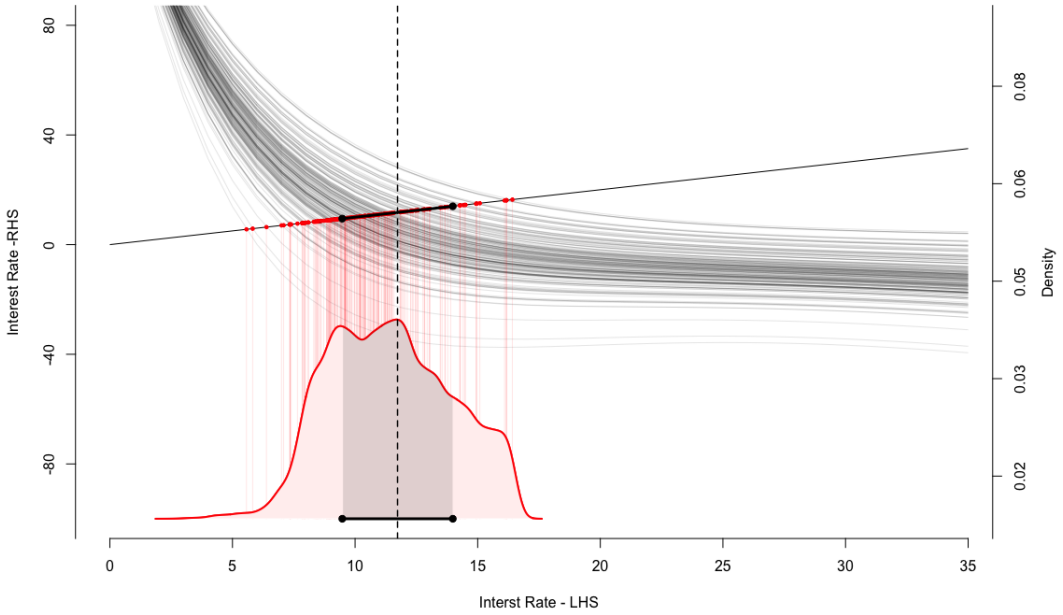
$$\frac{\partial \pi(r)}{\partial r} = L\left(r\dot{G}(r)\theta_R + G(r)\right)\left[1 - D(r)\right] - rG(r)\dot{D}(r)\delta_R = 0,$$

where $\dot{G}$ and $\dot{D}$ represent derivatives with respect to their scalar arguments, as before. Given the structural model, this simplifies to $(r(1 - G(r))\theta_R + 1) - rD(r)\delta_R = 0$. The optimal interest rate offer, denoted r^*, is therefore

$$r^* = \frac{1 + r^*(1 - G(r^*))\theta_R}{D(r^*)\delta_R}. \tag{5.3}$$

This is an implicit function but there will a unique fixed point since the numerator of the right hand side is decreasing in r while denominator is increasing, for $\theta_R < 0$ and $\delta_R > 0$. Figure 5 presents a visual representation of Equation (5.3). Each curve corresponds to a distinct consumer profile and the intersection with the 45^o line represents the fixed point r^*. The density then represents the kernel density of the optimal personalized offers (r^*) across consumers. We note that while the fixed

Figure 5: Optimal Personalized Interest Rate Offers



points are only shown for a subset of customers (to avoid clutter) the density is computed across the entire sample.

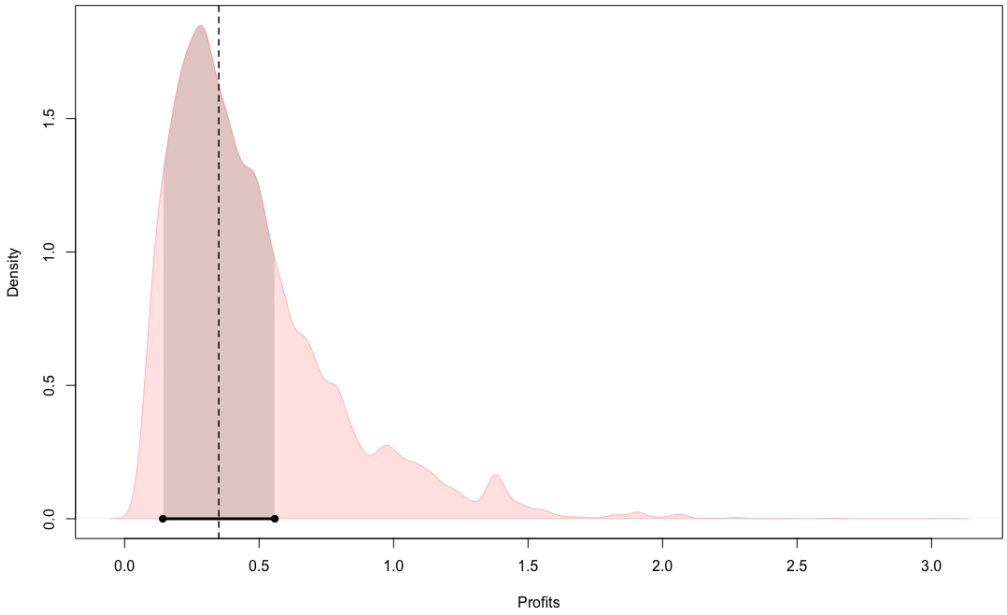
Even though r^* is not available in closed form it remains a smooth function of the parameters $\boldsymbol{\theta}$, which is all that is required for our method to apply. We can therefore give inference for any statistic of the form (2.2). As a simple example, Figure 5 shows estimation and inference for $\mu_0 = \mathbb{E}[r^*(\boldsymbol{\theta}(\boldsymbol{x}))]$, i.e. where the function H is the same as the function r^*. The vertical dashed line is the point estimate, found to be 11.37%, while a 95% confidence interval is shown as the grey region and black segment, founded to be $[9.48\%, 13.98\%]$.

Obtaining confidence intervals for more involved quantities is just as straightforward with our framework. We illustrate by computing the expected profits from setting the optimal personalized interest rate. From (5.2), this is expressed as

$$\mu_0 = \mathbb{E}\big[\pi(r^*(\boldsymbol{\theta}(\boldsymbol{X})))\big] = \mathbb{E}\Big[L\left[r^*(\boldsymbol{\theta}(\boldsymbol{X}))G(r^*(\boldsymbol{\theta}(\boldsymbol{X})))\right]\left[1 - D(r^*(\boldsymbol{\theta}(\boldsymbol{X})))\right]\Big].$$

Then, given the optimal interest rates r^*, and appealing to the envelope theorem we can obtain the influence function for this quantity ignoring the impact that perturbations in $\boldsymbol{\theta}$ have on r^* since

Figure 6: Expected Profits from Personalized Interest Rate Offers



$\frac{\partial \pi}{\partial r}|_{r=r^*} = 0$. As such, the influence function for expected profits can be constructed in closed form (conditional on r^*). Alternatively, one could use standard numerical differentiation or automatic differentiation engines to accomplish the same objective. In our analysis the numerical and exact derivatives give close to identical results. In our analysis we focus on the high risk segment (which is over 75% of the customers) and normalize the loan amount to $L = 1$. Since the interest rate is measured in percentage points, we interpret the expected profit construct μ_0 as the net average expected income from offering a \$100 loan at a personalized interest rate to each potential customer. We find that $\widehat{\mu}_0 = \$0.3504$ with a 95% confidence interval of $(\$0.1421, \$0.5586)$. Figure 6 depicts the density of profits for each customer along with the estimate and confidence interval of expected profits. The personalized interest rate scheme delivers an incremental 5.7% in expected profits over the optimal (uniform) interest rate derived from the experiment alone. While a more serious application would incorporate a number of additional features into the model and analysis, we feel that our example above suffices as a proof of concept of the value of our approach for applied work.

5.4 Summary

This application showcases the simplicity with which parametric models can be extended to incorporate nonparametric heterogeneity via deep neural networks. The structure of the model is maintained which in turn preserves the interpretability of the parameter functions. Since inference in our framework is close to automatic (automatic for data from randomized experiments) it offers the applied researcher a sophisticated yet practical framework for analysis.

6 Examples

Here we discuss several examples that fall within our framework, both to demonstrate the applicability of our results to new and interesting examples as well as to compare to existing results. We emphasize that these examples, and more, are covered without additional derivations: knowing the forms below is useful but not necessary before applying our methodology. This discussion is not exhaustive. We begin with two familiar examples, average treatment effects and partially linear models, before moving on to other cases.

6.1 Average Effect of a Binary Treatment

Average treatment effects are a canonical semiparametric problem and the standard case in the recent literature on inference after machine learning (see references in Section 4.1). Here we have a scalar outcome and $\boldsymbol{T} = T = \{0, 1\}$ is the scalar binary treatment indicator. The model is (2.3) with $G(u) = u$, so that $\mathbb{E}[Y \mid \boldsymbol{x}, t] = \theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x}) \cdot t$. Letting $Y(t)$ be the potential outcome under treatment $T = t$, we find that $\mathbb{E}[Y(0) \mid \boldsymbol{X} = \boldsymbol{x}] = \theta_{01}(\boldsymbol{x})$ and $\mathbb{E}[Y(1) \mid \boldsymbol{X} = \boldsymbol{x}] = \theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x})$, so that $\theta_{02}(\boldsymbol{x})$ represents the (heterogeneous) conditional average treatment effect, assuming unconfoundedness. Additional mean parameters could be added to cover average treatment effects for specific treatment groups as well as 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.

The naive approach to estimation would either involve unstructured modeling of $\mathbb{E}[Y \mid \boldsymbol{x}, t]$ or separate estimation (in the treatment and comparison groups) of $\mathbb{E}[Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]$ and $\mathbb{E}[Y(1) \mid \boldsymbol{X} = \boldsymbol{x}]$. Along with the propensity score, these would be inputs into the well-known doubly robust

or influence function estimator (Robins et al., 1994; Hahn, 1998). The structured architectures of Figures 1 and 2 intuitively reflect the idea that $\theta_{01}(\boldsymbol{x})$ and $\theta_{02}(\boldsymbol{x})$ may share similar features, since they relate to the conditional means of the two potential outcomes, under treatment and control. This same notion has been used in the past for trees by Zeileis et al. (2008) and Athey and Imbens (2016), where the treatment and control groups share a partition, and by Farrell et al. (2021) for DNNs, most similar to the present case. Notice that this is different from assuming that the regression functions share similar features to the propensity score, or its inverse, which in general there is no reason to expect to hold, particular in high-dimensional or data-adaptive scenarios. For classical, low-dimensional series estimators, this has been exploited to prove that both regression imputation (Imbens et al., 2007; Cattaneo and Farrell, 2011) and inverse weighting (Hirano et al., 2003) are semiparametrically efficient.

Equation (2.2) gives the familiar average treatment effect by taking $\boldsymbol{H}(\boldsymbol{x},\boldsymbol{\theta};\boldsymbol{t}^*)=\theta_{02}$. In this case, the model (2.1) is without loss of generality beyond unconfoundedness, and hence setting ℓ to be squared loss, we recover the familiar efficient influence function. To see this, begin with the univariate form in (4.5), and use the fact that $\dot{H}_1=0$, $\dot{H}_2=1$, $\lambda_0=1$, and $\lambda_1(\boldsymbol{x})=\lambda_2(\boldsymbol{x})=\mathbb{P}[T=1|\boldsymbol{X}=\boldsymbol{x}]:=p(\boldsymbol{x})$, the propensity score. Then, 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. To appreciate how simply and transparently our framework can be applied, suppose that beyond the mean effect, we were

interested in the variance of $Y(1)$ versus $Y(0)$. We take a quasi maximum likelihood approach, taking $\ell(\boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x}))$ to be the Gaussian likelihood, but instead of assuming constant variance (and thus fitting least squares regression) we optimization with respect to two additional parameters by taking the standard deviation to be $\sigma_1(\boldsymbol{x})t + \sigma_0(\boldsymbol{x})(1-t)$. The conditions for convexity of the loss are well-known from likelihood theory and can be directly used here.

6.2 Partially Linear Models

A second widely studied semiparametric problem is the partially linear model, where $G(\theta_{01}(\boldsymbol{x}) + \theta_{02}t)$, that is, where θ_{02} is assumed constant, and for simplicity we focus on a single scalar treatment variable. Restricting to a constant or homogeneous effect is a strong assumption and should be viewed with caution, but we can still apply our results to this case. Most studies use a linear model, but there are results for nonlinear $G(u)$. For our framework, we set the loss such that $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x})) = \boldsymbol{t}(G(\boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}) - y)$, as discussed above.

Typically, the parameter of interest is θ_{02}, in which case (4.5) gives that $\psi(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) - \theta_{02}$ is

$$\left[\lambda_2(\boldsymbol{x}) - \frac{\lambda_1(\boldsymbol{x})^2}{\lambda_0(\boldsymbol{x})}\right]^{-1} \left(t - \frac{\lambda_1(\boldsymbol{x})}{\lambda_0(\boldsymbol{x})}\right) \Big(y - G(\theta_{01}(\boldsymbol{x}) + \theta_{02}t)\Big).$$

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 received a great deal of attention in the literature, most often with a linear link function. Explicitly treating inference following machine learning, the pioneering work of Belloni et al. (2014) proved valid inference after lasso selection. Chernozhukov et al. (2018) use this model as the leading example of their generic results, and present several different Neyman orthogonal scores that could be used. Cattaneo et al. (2018) give novel results for series-based inference with many terms; they also give numerous references that use classical nonparametrics. For the case of nonlinear link function, Carroll et al. (1997) and Mammen and van de Geer (1997) study the nonparametric case, as we do here in Section 3, while Belloni et al. (2016) study high-dimensional

sparse models, where $\theta_{01}(\boldsymbol{x}) = \boldsymbol{\theta}_{01}'\boldsymbol{x}$. Our model is more general, but even so we obtain efficiency in the linear case under homoskedaticity, but not otherwise.

The literature has almost entirely focused on inference on the constant coefficient θ_{02}, but our framework allows for a much richer set of possibilities. For example, in both empirical finance and applied microeconomics the function $\theta_{01}(\boldsymbol{x})$ is of interest, see Cattaneo et al. (2020a) and Cattaneo et al. (2019) respectively.

6.3 Continuous Treatments and Average Partial Effects

Moving beyond discrete treatments or homogeneous effects, our framework gives a simple way to assess the heterogeneous effect of a continuous treatment T or set of treatments $\boldsymbol{T}$ by recovering average partial effects. In this case we begin with a linear model, $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \theta_{01} + \theta_{02}'\boldsymbol{t}$, and enrich the slopes and intercept to be parameter functions, so that $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \theta_{01}(\boldsymbol{x}) + \boldsymbol{\theta}_{02}(\boldsymbol{x})'\boldsymbol{t}$.

In this case, a common parameter is the average slope, or slopes, $\boldsymbol{\mu}_0 = \mathbb{E}[\boldsymbol{\theta}_{02}(\boldsymbol{x})]$. Although we are not restricted to this parameter, it is useful as it is the average of the heterogeneous partial effects, which, thanks to the model, can be extrapolated to any treatment level $\boldsymbol{t}^*$ by taking $\mathbb{E}[\boldsymbol{\theta}_{02}(\boldsymbol{x})'\boldsymbol{t}^*]$. Wooldridge (2004) and Graham and Pinto (2018) are the closest to our work in this example, and also give conditions for a causal interpretation of $\mathbb{E}[\theta_{0k}(\boldsymbol{X})]$. Hirshberg and Wager (2019) use a different approach to recover the average effect, but briefly discuss double robustness. Chernozhukov et al. (2019) use a similar model with the goal of policy targeting.

Our influence function specializes to exactly the efficient influence function of Graham and Pinto (2018) for $\boldsymbol{\mu}_0 = \mathbb{E}[\boldsymbol{\theta}_{02}(\boldsymbol{x})]$. Let $\mathbf{0}_d$ be the d-long zero vector, $\boldsymbol{I}_d$ be the d-square identity matrix, $\boldsymbol{V}(\boldsymbol{x}) = \mathbb{V}[\boldsymbol{T} \mid \boldsymbol{x}]$ be the conditional variance, and $\boldsymbol{E}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{T} \mid \boldsymbol{x}]$ be the conditional expectation. Then we have $\boldsymbol{\Lambda}(\boldsymbol{x}) = \begin{pmatrix} 1 & \boldsymbol{E}(\boldsymbol{x})' \\ \boldsymbol{E}(\boldsymbol{x}) & \mathbb{E}[\boldsymbol{T}\boldsymbol{T}'|\boldsymbol{x}] \end{pmatrix}$ and $\boldsymbol{H}_{\boldsymbol{\theta}} = (\mathbf{0}_{d_{\boldsymbol{\mu}}}\ ,\ \boldsymbol{I}_{d_{\boldsymbol{\mu}}})$, so after some algebra, Equation (4.4) gives

$$\boldsymbol{\psi}(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = \boldsymbol{\theta}_{02}(\boldsymbol{x}) - \boldsymbol{V}(\boldsymbol{x})^{-1}(\boldsymbol{t} - \boldsymbol{E}(\boldsymbol{x}))(y - \theta_{01}(\boldsymbol{x}) - \boldsymbol{\theta}_{02}(\boldsymbol{x})'\boldsymbol{t}).$$

A simple, but useful extension to the scalar case is when the vector $\boldsymbol{T}$ includes polynomials, or other flexible specifications, or interaction terms of several policy variables, and we wish to study heterogeneous effects. That is, for two treatment variables T_1 and T_2, we may be interested in the coefficients on T_1^2 or $T_1 \times T_2$. Taking T_2 to be a binary or categorical variable would yield subgroup

effects. Such objects are routinely studied in the parametric case, and here we allow full heterogeneity in these effects, beyond the original treatment or partial effect.

The example of average partial effects, when combined with transformations of the outcome and the treatment, recovers many other useful settings, even restricting to linear link functions. To give just two examples, consider the so-called Berry logit and Cobb-Doublas production. The former, pioneered by Berry (1994), is a popular model for demand models where the outcome of interest is the market share distribution across firms. In most applications the researcher has access to outcomes $\{Y_{jm}\}$ 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

$$\mathbb{E}\left[\log \frac{Y_{jm}}{Y_{0m}} \middle| \boldsymbol{X} = \boldsymbol{x}_m, \boldsymbol{T}_{jm} = \boldsymbol{t}_{jm}\right] = \theta_{01j}(\boldsymbol{x}_m) + \boldsymbol{\theta}_{02}(\boldsymbol{x}_m)'(\boldsymbol{t}_{jm} - \boldsymbol{t}_{0m}).$$

Stacking these equations and the corresponding data allows us to construct an estimator for $\theta_{01j}(\boldsymbol{x}_m)$ and $\boldsymbol{\theta}_{02}(\boldsymbol{x}_m)$. We note here that our framework can be extended to include instruments along the lines of Okui et al. (2012).

The Cobb-Douglas specification for production functions is denoted by $Y = CK^{\theta_{01}(\boldsymbol{x})}L^{\theta_{02}(\boldsymbol{x})}$. With $\boldsymbol{T} = (K, L)'$, by taking logs we can write this model in our format, as

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

Given estimates we may be interested in understanding if the technology exhibits increasing, constant, or decreasing returns to scale. This can be ascertained by computing $\boldsymbol{\mu}_0 = \mathbb{E}[\theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x})]$ and noting that $\boldsymbol{\mu}_0 < 1, \boldsymbol{\mu}_0 = 1, \boldsymbol{\mu}_0 > 1$ imply decreasing, constant, and increasing returns to scale. 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.

It is useful to contrast our model, $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \theta_{01}(\boldsymbol{x}) + \boldsymbol{\theta}_{02}(\boldsymbol{x})'\boldsymbol{t}$, with the fully unrestricted case, $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \theta_0(\boldsymbol{x}, \boldsymbol{t})$. For causal inference in particular, this case has been studied by Hirano and

Imbens (2004) and, using doubly robust approaches, 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. Here our model imposes nontrivial structure, unlike in the binary case, but yields a tractable and interpretable model. From a practical point of view, compared to $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \theta_0(\boldsymbol{x}, \boldsymbol{t})$, our approach results in lower dimensional estimation and does not require conditional density estimation, which can be challenging in high dimensional, complex settings.

Another related area is the study of (weighted) average derivatives, a common estimand in the literature on semiparametric theory (Powell et al., 1989; Newey and Stoker, 1993). Here the object of interest is $\mathbb{E}[w(\boldsymbol{x}, \boldsymbol{t})\partial\theta(\boldsymbol{x}, \boldsymbol{t})/\partial\boldsymbol{t}]$ for a known weighting function $w(\boldsymbol{x}, \boldsymbol{t})$. 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.

6.4 Fractional Outcomes

Building on the previous example, we emphasize that nonlinear models be covered seamlessly, given appropriate regularity. One widely-used case is fractional outcome models, following Papke and Wooldridge (1996). In these models the outcome Y is continuous but restricted to lie in $[0, 1]$. In that paper, the sampling units are firms and Y is the rate of employee partitioning in 401(k) plans. The policy variable is the employer match rates. Papke and Wooldridge (1996) apply a parametric logitistic QLME, explicitly advocating the use of structure to ensure that the outcomes remain on the unit interval. They argue that this specification is valid even at the endpoints and is more practically relevant then transformations of the dependent variable.

We take their model and enrich it to allow for heterogeneity. In our notation, Papke and Wooldridge (1996) assume $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, T = t] = G(\theta_{01} + \theta_{02} \cdot t + \boldsymbol{\gamma}'\boldsymbol{x})$, with G the logistic link; a structured, but rigid, parametric model with the covariates. We allow for the much more general $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, T = t] = G(\theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x}) \cdot t)$.

The substantive quantities of interest in the original application is the marginal effect of the match rate on participation and the degree to which this marginal effect exhibits diminishing patterns. To investigate this Papke and Wooldridge (1996) evaluate the marginal effect at fixed values of $\mathbf{x}$ and several match rates, $t^* \in \{0.0, 0.5, 1.0\}$. They conclude that there exists evidence for diminishing

marginal effects, for example.

We can generalize these findings by conducting inference on the average marginal effect (AME) and the average change in the marginal effect (ACME[5]), given by

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

Because of the structure of the model, these are easily recovered in the form of $\boldsymbol{\mu}_0$, by taking $H_{\text{AME}}\,(\boldsymbol{x}, \boldsymbol{\theta}; t^*) = \theta_{02} G^*\,(1 - G^*)$ and $H_{\text{ACME}}\,(\boldsymbol{x}, \boldsymbol{\theta}; t^*) = \theta_{02}^2 G^*\,(1 - G^*)\,(1 - 2G^*)$, respectively, where $G^* = G(\boldsymbol{\theta}'\boldsymbol{t}^*)$. Note the contrast with the naive, unstructured ML approach, where recovering the second derivative of a complex, high-dimensional $G(\hat{u}(\boldsymbol{x}, t))$ could be challenging.

Our deep neural networks can be structured to respect fractional losses by using the QMLE logistic loss as the model (2.1) while assuming the mean restriction (2.3). For estimation and inference, we must assume that $\mathbb{E}[\dot{G}\boldsymbol{T}\boldsymbol{T}'|\boldsymbol{x}]$ is nonsingular, again a weighted conditional variance assumption. Theorem 2 (or in this case, Equations (4.4) or (4.5)) apply immediately. The derivatives required for ℓ are well known from likelihood theory and can be used directly. Those for H are easily available or can be computed if necessary.

6.5 Type I Tobit

To illustrate the use of existing likelihood theory for parametric models and how these can help interpret our requirements, consider the type I Tobit model. This is a case where our assumptions are conditional versions of the standard conditions required for parametric MLE and therefore we can intuitively understand our conditions by imagining our method as if it was parametric MLE for each value $\boldsymbol{x}$.

Here we assume that the observed outcome is $Y = \max(0, Y^*)$, where Y^* is Gaussian given $\boldsymbol{x}$ with mean given by, say $\boldsymbol{\beta}_0(\boldsymbol{x})'\boldsymbol{t}$ and variance $\sigma^2(\boldsymbol{x})$ (in practice one may take $\sigma^2(\boldsymbol{x}) = \exp\{\tilde{\sigma}(\boldsymbol{x})\}$ for example). In this case, we work with the transformed parameters $\boldsymbol{\theta}(\boldsymbol{x}) = (\boldsymbol{\theta}_1(\boldsymbol{x})', \theta_2(\boldsymbol{x}))'$, with $\boldsymbol{\theta}_1(\boldsymbol{x}) = \boldsymbol{\beta}_0(\boldsymbol{x})/\sigma(\boldsymbol{x})$ and $\theta_2(\boldsymbol{x}) = \sigma^{-1}(\boldsymbol{x})$. See Amemiya (1985) and Wooldridge (2010) for textbook treatments and details on the calculations below.

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

[5] We apologize for the acronym. Alternative suggestions for this effect are welcome.

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. Then the gradient (score) terms are

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

and

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

The second derivatives are

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

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

That the gradients are conditionally mean zero can be directly verified. The matrix $\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}$ exists because $\boldsymbol{\theta}_1(\boldsymbol{x})'\boldsymbol{t} - \phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\boldsymbol{t})/[1 - \Phi(\boldsymbol{\theta}_1(\boldsymbol{x})'\boldsymbol{t})] > 0$, using exactly the logic from parametric models (Donald, 1990; Olsen, 1978; Amemiya, 1985). Naturally other conditions, such as smoothness, would be required for the functions $\boldsymbol{\beta}(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ and would need to be matched by the neural network, or other nonparametric estimator.

6.6 Multinomial Choice

The binary choice model examined in the application naturally extends to multiple choices. Here the model (2.1) deals with a vector of outcomes. Let there be $J \geq 1$ choices, in addition to the outside option. The outcome is the categorical variable $Y \in \{0, 1, \ldots, J\}$ or the vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_J)'$, with $Y_j = \mathbb{1}\{Y = j\}$. The standard assumption is that

$$\mathbb{P}[Y = j \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G_j\left(u_1, u_2, \ldots, u_J\right), \qquad \text{with} \qquad G_j = \frac{\exp\{u_j\}}{1 + \sum_{m=1}^{J} \exp\{u_m\}},$$

for utility functions $u_j = u_j(\boldsymbol{x}, \boldsymbol{t})$, with u_0 normalized to zero. If we let $G_j = G_j(u_0, u_1, \ldots, u_J)$, then using that $Y_0 + Y_1 + \cdots + Y_J = 1$, $u_0 = 0$, and the form of G_j, the log-likelihood is

$$\ell = \log(G_0) + \sum_{j=1}^{J} y_j u_j.$$

The negative would be the loss we minimize.

The specification of the utility functions gives rise to different parametric models in this context. The defacto standard in many disciplines is McFadden's multinomial choice model, and our enriched version would assume that the utilities obey

$$u_j(\boldsymbol{x}, \boldsymbol{t}_j) = \theta_{01j}(\boldsymbol{x}) + \boldsymbol{\theta}_{02}(\boldsymbol{x})' \boldsymbol{t}_j, \quad j = 1, \ldots, J,$$

where the $\boldsymbol{t}_j$ are option-specific characteristics, such as prices. The key restriction is that, while the intercept functions are choice-specific, the price effect functions are common across options. This model is well studied, and the gradient $\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x}))$ and Hessian $\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w}, \boldsymbol{\theta}(\boldsymbol{x}))$ are well understood. The parameter $\boldsymbol{\mu}_0$ could depend on any of the intercept and slope functions. The orthogonal score can then be read off from Theorem 2. The result is completely new to the literature and, as above, fully implementable.

6.7 Instrumental Variables

For linear models, such as in Section 6.1 or 6.3, a natural extension is to endogenous variables of interest. Consider for simplicity the case with a single endogenous treatment variable and a single instrument Z. It is natural in our setting to allow fully flexible observed heterogeneity in the effects of the instrument. We therefore arrive at the two-equation model

$$Y = \theta_{01}(\boldsymbol{X}) + \theta_{02}(\boldsymbol{X}) T + V, \tag{6.1}$$

$$T = \zeta_{01}(\boldsymbol{X}) + \zeta_{02}(\boldsymbol{X}) Z + U, \tag{6.2}$$

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 (6.2) into (6.1) to obtain the reduced form equation

$$\begin{aligned} &Y = \alpha_0(\boldsymbol{X}) + \beta_0(\boldsymbol{X})Z + \tilde{V}, \\ &\alpha_0(\boldsymbol{x}) = \theta_{01}(\boldsymbol{x}) + \theta_{02}(\boldsymbol{x})\zeta_{01}(\boldsymbol{x}), \quad \beta_0(\boldsymbol{x}) = \theta_{02}(\boldsymbol{x})\zeta_{02}(\boldsymbol{x}), \quad \tilde{V} = \theta_{02}(\boldsymbol{X})U + V. \end{aligned} \tag{6.3}$$

Using the instruments in this way directly generalizes the standard two stage least squares approach to handle high-dimensional, complex observed heterogeneity. Deep learning is again well-suited to estimating the coefficient functions in (6.2) and (6.3), exactly following Section 3. The loss (2.1) is simply the sum of the two squared losses.

With this notation, we aim to recover a parameter that depends on the coefficient functions of (6.2) and (6.3), given by

$$\boldsymbol{\mu}_0 = \mathbb{E}\left[H\left(\boldsymbol{X}, \alpha_0, \beta_0, \zeta_{01}, \zeta_{02}; \boldsymbol{t}^*\right)\right]. \tag{6.4}$$

The leading case is the average partial effect of the endogenous variable T: $\boldsymbol{\mu}_0 = \mathbb{E}[\theta_{02}(\boldsymbol{X})] = \mathbb{E}[\beta_0(\boldsymbol{X})/\zeta_{02}(\boldsymbol{X})]$. Note that here we are assuming the analogue of strong instruments, as we need $\zeta_{02}(\boldsymbol{X})$ to be nowhere zero.

To show the score in this case, define $\boldsymbol{\theta} = (\alpha_0, \beta_0, \zeta_{01}, \zeta_{02})'$, $\boldsymbol{w} = (y, t, z)$, $\boldsymbol{t} = (1, t)'$, $\boldsymbol{z} = (1, z)'$ and $\boldsymbol{I}_2$ the 2×2 identity. Then we have

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

Therefore $\boldsymbol{\Lambda}(\boldsymbol{x}) = \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_Z(\boldsymbol{x})$, where $\boldsymbol{\Lambda}_Z(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}' \mid \boldsymbol{X} = \boldsymbol{x}]$. These can be inserted directly into Equation (4.3).

This approach is far from the only option in instrumental variable models. Indeed, for the special case of homogeneous effects in a partially linear IV model, Chernozhukov et al. (2018) study two different scores, Equations (4.7) and (4.8) therein, and also mention in footnote 8 that their stated method for constructing orthogonal scores would yield a third option. Different scores sometimes require that different functions be estimated in the first step or as part of the correction term. Our approach here aims for ease of use and transparency: (6.2) and (6.3) can be directly estimated using the deep learning architectures of Section 3.

7 Extensions

Our framework is connected to many different areas of debiasing for inference, some of which have been explored in settings we have ruled out. Our methodological ideas could be extended to many of these settings.

Building on our two-step approach, we could consider more general two-step GMM type problems, where the first step has been enriched with deep learning. In some cases, our results can be extended directly, at mainly a notational cost. For example, $\boldsymbol{\mu}_0$ need not be restricted to a closed form, but could defined in terms of an objective function or moment condition itself. That is, (2.2) could be changed such that $\boldsymbol{\mu}_0$ solved $\max_{\boldsymbol{\mu}} \mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)]$ or $\mathbb{E}[\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\theta}_0(\boldsymbol{X}); \boldsymbol{t}^*)] = 0$. In other cases, more substantial work would be necessary. Extending our ideas to quantile regression and other nonsmooth objective functions would be such an example. The extent to which general, easily implementable results can be given in such cases is an interesting avenue for future research.

An important point of our framework is that the parameter functions $\boldsymbol{\theta}(\boldsymbol{x})$ have economic meaning and interpretability. A useful extension to this would be to consider shape restrictions; see Chetverikov et al. (2018) for review. For example, price coefficients should be nonpositive, and $\theta < 0$ is often found or enforced in parametric modeling. In our context, we would like to ensure that $\theta(\boldsymbol{x}) < 0$. In our experience, the discipline of the model often yields functions which empirically obey such restrictions, i.e., loosely speaking, the model regularizes the data toward economically-valid estimates. However, this does not always hold, and it would be interested to enforce this during the estimation and establish second-step inference. One possibility to enforce such shape constraints is by designing a proper barrier function added to the original loss, thus to leverage techniques developed in constrained optimization (Nesterov and Nemirovskii, 1994). Provided the approximation results still hold, a version of Theorem 1 could be obtained, after which inference using the influence function can proceed.

Another important extension for some applications would be to consider the case when the number of variables of interest is large. In this case, our ideas connect with the literature on debiasing in high dimensional regression. Consider the case of a conditional mean restriction with a linear function, so that $\mathbb{E}[Y \mid \boldsymbol{x}, \boldsymbol{t}] = \boldsymbol{\theta}_0(\boldsymbol{x})'\boldsymbol{t}$. A fundamental tension exists between the dimensionality of $\boldsymbol{T}$ and the complexity allowed for in $\boldsymbol{\theta}_0(\boldsymbol{X})$. We have studied the case of fully flexible heteregeneity

for a low-dimensional $\boldsymbol{T}$. When $\boldsymbol{T}$ is high dimensional, estimation of these functions will not be possible, at least not with sufficient precision to allow for inference. There is thus a natural tradeoff between the dimensionality of the treatment variables and the complexity of the heterogeneity. If a researcher can a priori restrict the form of the heterogeneity such that the functions $\boldsymbol{\theta}_0(\boldsymbol{x})$ are simple, or in the extreme case, constant, then useful results can be obtained. For this model, Javanmard and Montanari (2018) seek a Gaussian limit for $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ where

$$\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{\text{lasso}} + \frac{1}{n}\widehat{\boldsymbol{\Lambda}}\boldsymbol{T}_n\left(\boldsymbol{Y}_n - \boldsymbol{T}_n\widehat{\boldsymbol{\theta}}_{\text{lasso}}\right),$$

where $\widehat{\boldsymbol{\theta}}_{\text{lasso}}$ is the lasso estimator and $\widehat{\boldsymbol{\Lambda}}$ is an estimator of $\boldsymbol{\Lambda} = \mathbb{E}[\boldsymbol{T}\boldsymbol{T}']$, which is not a function of $\boldsymbol{X}$ in this restricted model, and the data is $\boldsymbol{Y}_n = (y_1, \ldots, y_n)$ and $\boldsymbol{T}_n = (\boldsymbol{t}_1', \ldots, \boldsymbol{t}_n')'$. The above display is in perfect analogy with our Theorem 2 (in particular, the form in Corollary 2), and the second term serves essentially the same function in both cases. One may check that perturbations to $\boldsymbol{\theta}$ do not have a first order impact in expectation, as required for Neyman orthogonality. Similar to our Theorem 3, they require that $\widehat{\boldsymbol{\theta}}_{\text{lasso}}$ and $\widehat{\boldsymbol{\Lambda}}$ are "good enough" first-stage estimators of $\boldsymbol{\theta}_0$ and $\boldsymbol{\Lambda}$, respectively, which they prove for sparse regression under conditions on the design $\boldsymbol{T}_n$.

The linear model above rules out all heterogeneity, and is therefore less interesting for our present purpose. An open question is how much flexible heterogeneity can be accommodated which still obtaining useful results. Moving slightly beyond constant effects, Kozbur (2020) allows for a flexible intercept, so that $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = \theta_{01}(\boldsymbol{x}) + \boldsymbol{\theta}_{02}'\boldsymbol{t}$, and studies inference on functionals. Our ideas can be adapted to this model. The architecture we proposed in Section 3, shown in Figure 2, will need adjustment. Instead of learning the functions in the parameter layer, the model layer will learn weights for edge between $\boldsymbol{t}$ and the outcome. Some form of regularization will be needed in this layer, and thus the end result will be a combination of deep learning and regularized high dimensional regression. A formal exploration of this is a promising direction for future research.

8 Conclusion

We have provided a complete methodological framework for using machine learning to enrich economic models to exploit rich, complex data on individual heterogeneity. We showed that deep learning

is ideally suited to this task among modern machine learning methods and we detailed a new network architecture that is designed to estimate economically meaningful objects, moving past pure prediction and towards structural modeling. We gave results for the estimation of heterogeneity using deep learning, showing how our architecture delivers improved rates of convergence. Subsequent inference is proven valid building on a newly calculated influence function with broad applicability.

Our framework covers a wide variety of interesting contexts. The combination of the specification we adopt, the availability of computing infrastructure, and the theory presented above offer a perfect package for applied researchers.

9 References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng (2015): "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," Software available from tensorflow.org. (Cited on page 29.)

Ackerberg, D., X. Chen, and J. Hahn (2012): "A practical asymptotic variance estimator for two-step semiparametric estimators," *Review of Economics and Statistics*, 94, 481–498. (Cited on page 23.)

Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press. (Cited on pages 42 and 43.)

Athey, S. and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360. (Cited on page 37.)

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178. (Cited on page 11.)

Babii, A., X. Chen, E. Ghysels, and R. Kumar (2020): "Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice," *arXiv:2010.08463*. (Cited on page 12.)

Bach, F. (2017): "Breaking the curse of dimensionality with convex neural networks," *The Journal of Machine Learning Research*, 18, 629–681. (Cited on page 17.)

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005): "Local rademacher complexities," *The Annals of Statistics*, 33, 1497–1537. (Cited on page 56.)

Bartlett, P. L., N. Harvey, C. Liaw, and A. 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 58.)

Bauer, B. and M. Kohler (2019): "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Annals of Statistics*, 47, 2261–2285. (Cited on page 17.)

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650. (Cited on pages 18, 19, and 38.)

Belloni, A., V. Chernozhukov, and Y. Wei (2016): "Post-selection inference for generalized linear models with many controls," *Journal of Business & Economic Statistics*, 34, 606–619. (Cited on page 38.)

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, 25, 242–262. (Cited on page 40.)

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman (2010): "What's advertising content worth? Evidence from a consumer credit marketing field experiment," *The Quarterly Journal of Economics*, 125, 263–306. (Cited on pages 1, 28, 29, and 30.)

Blanchet, J., Y. Kang, J. L. M. Olea, V. A. Nguyen, and X. Zhang (2020): "Machine Learning's Dropout Training is Distributionally Robust Optimal," *arXiv:2009.06111*. (Cited on page 17.)

Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997): "Generalized partially linear single-index models," *Journal of the American Statistical Association*, 92, 477–489. (Cited on page 38.)

Cattaneo, M. D. (2010): "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155, 138–154. (Cited on page 36.)

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2019): "On Binscatter," *arXiv:1902.09608*. (Cited on page 39.)

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and E. Schaumburg (2020a): "Characteristic-Sorted Portfolios: Estimation and Inference," *Review of Economics and Statistics*, 101, 531–551. (Cited on page 39.)

Cattaneo, M. D. and M. H. Farrell (2011): "Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates," in *Advances in Econometrics: Missing Data Methods*, ed. by D. Drukker, Emerald Group Publishing Limited, vol. 27A, 93–127. (Cited on pages 36 and 37.)

——— (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174, 127–143. (Cited on page 11.)

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2020b): "Large Sample Properties of Partitioning-based Series Estimators," *Annals of Statistics*, 48, 1718–1741. (Cited on page 11.)

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity," *Journal of the American Statistical Association*, 113, 1350–1361. (Cited on page 38.)

Chatla, S. B. and G. Shmueli (2020): "A Tree-Based Semi-Varying Coefficient Model for the COM-Poisson Distribution," *Journal of Computational and Graphical Statistics*, 29, 827–846. (Cited on page 11.)

Chen, H., A. Didisheim, and S. Scheidegger (2021): "Deep Structural Estimation: With an Application to Option Pricing," *arXiv preprint:2102.09209*. (Cited on page 12.)

Chen, R. and R. Tsay (1993): "Functional-coefficient autoregressive models," *Journal of the American Statistical Association*, 88, 298–308. (Cited on page 9.)

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 6B of *Handbook of Econometrics*, chap. 76. (Cited on pages 3, 12, and 13.)

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68. (Cited on pages 3, 18, 19, 25, 26, 27, 38, 45, and 63.)

Chernozhukov, V., M. Demirer, G. Lewis, and V. Syrgkanis (2019): "Semi-Parametric Efficient Policy Learning with Continuous Actions," in *Advances in Neural Information Processing Systems 32*, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, 15065–15075. (Cited on page 39.)

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2020a): "Locally Robust Semiparametric Estimation," *arXiv:1608.00033*. (Cited on pages 18 and 19.)

Chernozhukov, V., W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis (2021): "Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression," *arXiv:2104.14737*. (Cited on page 18.)

Chernozhukov, V., W. K. Newey, and R. Singh (2020b): "Automatic Debiased Machine Learning of Causal and Structural Effects," *arXiv:1809.05224*. (Cited on page 18.)

——— (2020c): "De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers," *arXiv:1802.08667*. (Cited on page 18.)

Chernozhukov, V., W. K. Newey, R. Singh, and V. Syrgkanis (2020d): "Adversarial Estimation of Riesz Representers," *arXiv preprint arXiv:2101.00009*. (Cited on page 18.)

Chetverikov, D., A. Santos, and A. M. Shaikh (2018): "The Econometrics of Shape Restrictions," *Annual Review of Economics*, 10, 31–63. (Cited on page 46.)

Cleveland, W. S., E. Grosse, and W. M. Shyu (1991): "Local regression models," in *Statistical models in S*, ed. by J. M. Chambers and T. Hastie, Pacific Grove: Wadsworth and Brooks/Cole, 309–376. (Cited on page 9.)

Colangelo, K. and Y.-Y. Lee (2020): "Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments," *arXiv:2004.03036*. (Cited on page 41.)

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008): "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90, 389–405. (Cited on page 22.)

Donald, S. G. (1990): "Estimation of heteroskedastic limited dependent variable models," Ph.D. thesis, University of British Columbia. (Cited on page 43.)

Fan, J. and W. Zhang (2008): "Statistical methods with varying coefficient models," *Statistics and Its Interface*, 1, 179–195. (Cited on page 11.)

Farrell, M. 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 18 and 36.)

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep Neural Networks for Estimation and Inference," *arXiv:1809.09953, Econometrica*, 89, 181–213. (Cited on pages 3, 10, 13, 14, 17, 26, 27, 37, 56, 57, and 58.)

Foster, D. J. and V. Syrgkanis (2020): "Orthogonal statistical learning," *arXiv preprint arXiv:1901.09036*. (Cited on page 24.)

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep learning*, Cambridge: MIT Press. (Cited on page 10.)

Graham, B. S. and C. C. d. X. Pinto (2018): "Semiparametrically efficient estimation of the average linear regression function," *Journal of Econometrics,* forthcoming. (Cited on page 39.)

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. (Cited on pages 18 and 37.)

——— (2004): "Functional restriction and efficiency in causal inference," *Review of Economics and Statistics*, 84, 73–76. (Cited on page 18.)

Hanin, B. (2017): "Universal function approximation by deep neural nets with bounded width and relu activations," *arXiv preprint arXiv:1708.02691*. (Cited on page 13.)

Hastie, T. and R. Tibshirani (1993): "Varying-Coefficient Models," *Journal of the Royal Statistical Society, Series B*, 55, 757–796. (Cited on page 9.)

Hirano, K. and G. W. Imbens (2004): "The Propensity Score with Continuous Treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. by G. A. and X.-L. Meng, New York: Wiley, 73–84. (Cited on page 40.)

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. (Cited on pages 18 and 37.)

Hirshberg, D. A. and S. Wager (2019): "Augmented Minimax Linear Estimation," *arXiv:1712.00038*. (Cited on page 39.)

Huang, J. Z. and H. Shen (2004): "Functional coefficient regression models for non-linear time series: a polynomial spline approach," *Scandinavian Journal of Statistics*, 31, 515–534. (Cited on page 59.)

Ichimura, H. and W. K. Newey (2015): "The influence function of semiparametric estimators," *arXiv preprint arXiv:1508.01378*. (Cited on pages 18, 19, and 59.)

Igami, M. (2020): "Artificial intelligence as structural estimation: Deep Blue, Bonanza, and AlphaGo," *The Econometrics Journal,* forthcoming. (Cited on page 12.)

Imbens, G. W., W. K. Newey, and G. Ridder (2007): "Mean-Squared-Error Calculations for Average Treatment Effects," *working paper*. (Cited on pages 18 and 37.)

Javanmard, A. and A. Montanari (2018): "Debiasing the lasso: Optimal sample size for gaussian designs," *The Annals of Statistics*, 46, 2593–2622. (Cited on page 47.)

Kaji, T., E. Manresa, and G. Pouliot (2020): "An adversarial approach to structural estimation," *arXiv preprint arXiv:2007.06169*. (Cited on page 12.)

Kennedy, E. H. (2020): "Optimal doubly robust estimation of heterogeneous causal effects," *arXiv:2004.14497*. (Cited on page 24.)

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 1229. (Cited on page 41.)

Kingma, D. P. and J. Ba (2014): "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*. (Cited on page 29.)

Kozbur, D. (2020): "Inference in additively separable models with a high-dimensional set of conditioning variables," *Journal of Business & Economic Statistics,* forthcoming. (Cited on page 47.)

Li, Q., C. J. Huang, D. Li, and T.-T. Fu (2002): "Semiparametric smooth coefficient models," *Journal of Business & Economic Statistics*, 20, 412–422. (Cited on page 9.)

Liang, T. (2018): "On How Well Generative Adversarial Networks Learn Densities: Nonparametric and Parametric Results," *arXiv:1811.03179*. (Cited on page 17.)

Liang, T. and H. Tran-Bach (2020): "Mehler's Formula, Branching Process, and Compositional Kernels of Deep Neural Networks," *arXiv preprint arXiv:2004.04767*. (Cited on page 17.)

Ma, X. and J. Wang (2020): "Robust inference using inverse probability weighting," *Journal of the American Statistical Association*, 115, 1851–1860. (Cited on page 22.)

Mammen, E. and S. van de Geer (1997): "Penalized quasi-likelihod estimation in partially linear models," *Annals of Statistics*, 25, 1014–1035. (Cited on pages 23 and 38.)

Maurer, A. (2016): "A Vector-Contraction Inequality for Rademacher Complexities," in *Algorithmic Learning Theory*, ed. by R. Ortner, H. U. Simon, and S. Zilles, Cham: Springer International Publishing, 3–17. (Cited on page 57.)

Nekipelov, D., V. Semenova, and V. Syrgkanis (2020): "Regularized Orthogonal Machine Learning for Nonlinear Semiparametric Models," *arXiv:1806.04823*. (Cited on page 24.)

Nesterov, Y. and A. Nemirovskii (1994): *Interior-point polynomial algorithms in convex programming*, SIAM. (Cited on page 46.)

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–342. (Cited on page 40.)

Newey, W. K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135. (Cited on page 59.)

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. (Cited on pages 3, 18, 21, and 59.)

Newey, W. K. and D. L. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. McFadden, Elsevier, vol. 4 of *Handbook of Econometrics*, chap. 36, 2111–2245. (Cited on pages 8, 19, and 23.)

Newey, W. K. and J. M. Robins (2018): "Cross-fitting and fast remainder rates for semiparametric estimation," *arXiv preprint arXiv:1801.09138*. (Cited on page 26.)

Newey, W. K. and T. M. Stoker (1993): "Efficiency of weighted average derivative estimators and index models," *Econometrica*, 61, 1199–1223. (Cited on page 41.)

Nie, X. and S. Wager (2020): "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," *arXiv:1712.04912. Biometria,* forthcoming. (Cited on page 24.)

O'Hagan, A. (1978): "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society: Series B*, 40, 1–24. (Cited on page 9.)

Okui, R., D. S. Small, Z. Tan, and J. M. Robins (2012): "Doubly Robust Instrumental Variable Regression," *Statistica Sinica*, 22, 173–205. (Cited on page 40.)

Olsen, R. J. (1978): "Note on the uniqueness of the maximum likelihood estimator for the Tobit model," *Econometrica*, 46, 1211–1215. (Cited on page 43.)

Padilla, O. H. M., W. Tansey, and Y. Chen (2020): "Quantile regression with ReLU Networks: Estimators and minimax rates," *arXiv preprint arXiv:2010.08236*. (Cited on page 13.)

Papke, L. E. and J. M. Wooldridge (1996): "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates," *Journal of Applied Econometrics*, 11, 619–632. (Cited on page 41.)

Polson, N. G. and V. Ročková (2018): "Posterior concentration for sparse deep learning," in *Advances in Neural Information Processing Systems*, 930–941. (Cited on page 17.)

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. (Cited on page 41.)

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. (Cited on page 37.)

Schmidt-Hieber, J. (2019): "Nonparametric regression using deep neural networks with ReLU activation function," *arXiv:1708.06633, Annals of Statistics*, forthcoming. (Cited on page 17.)

Stone, C. J., M. H. Hansen, C. Kooperberg, and Y. K. Truong (1997): "Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture," *The Annals of Statistics*, 25, 1371–1470. (Cited on page 9.)

Tan, Z. (2020): "Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data," *Annals of Statistics*, 48, 811–837. (Cited on page 24.)

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press. (Cited on page 59.)

Wang, Y. and V. Ročková (2020): "Uncertainty Quantification for Sparse Deep Learning," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. (Cited on page 17.)

Wei, Y. and Z. Jiang (2019): "Estimating Parameters of Structural Models Using Neural Networks," *SSRN 3496098*. (Cited on page 12.)

Wooldridge, J. M. (2004): "Estimating average partial effects under conditional moment independence assumptions," *cemmap working paper CWP03/04*. (Cited on page 39.)

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press, 2 ed. (Cited on page 42.)

Yarotsky, D. (2017): "Error bounds for approximations with deep ReLU networks," *Neural Networks*, 94, 103–114. (Cited on pages 13 and 58.)

——— (2018): "Optimal approximation of continuous functions by very deep ReLU networks," *arXiv preprint arXiv:1802.03620*. (Cited on page 13.)

Zeileis, A., T. Hothorn, and K. Hornik (2008): "Model-based recursive partitioning," *Journal of Computational and Graphical Statistics*, 17, 492–514. (Cited on pages 11 and 37.)

A Proofs for Deep Learning

We now prove results stated in the main text for deep learning. First, we show that $\boldsymbol{\theta}_0(\boldsymbol{x})$ is identified and then we prove the rates of convergence for these functions. Lastly we discuss estimation of $\boldsymbol{\Lambda}(\boldsymbol{x})$.

A.1 Proof of Theorem 1

The proof method of Farrell et al. (2021) is used here. Some details will be deferred to that paper.

Let M be such that $\max_{k\leq d_{\boldsymbol{\theta}}} \|\theta_{0k}\|_\infty < M$ and functions computed by $\mathcal{F}_{\text{DNN}}$ are similarly bounded by $2M$. Define $\boldsymbol{\theta}_n \in \mathcal{F}_{\text{DNN}}$ as the best approximation to $\boldsymbol{\theta}_0$ in the class of DNNs and let ϵ_n denote the error of the approximation:

$$\boldsymbol{\theta}_n = \underset{\substack{\boldsymbol{\theta}\in\mathcal{F}_{\text{DNN}} \\ \|\boldsymbol{\theta}\|_\infty \leq 2M}}{\arg\min} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty, \qquad \epsilon_n = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|_\infty.$$

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. Here we allow for other approximation assumptions (such as other smoothness classes) and other architectures by leaving the approximation generic.

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

$$\begin{aligned}
&c_1\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] \\
&\quad\leq \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))] - \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))] \\
&\quad\leq \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))] - \mathbb{E}[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))] - \mathbb{E}_n[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))] + \mathbb{E}_n[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_n(\boldsymbol{X}))] \\
&\quad= (\mathbb{E}-\mathbb{E}_n)\left[\ell(\boldsymbol{Y},\boldsymbol{T},\widehat{\boldsymbol{\theta}}(\boldsymbol{X})) - \ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))\right] + \mathbb{E}_n\left[\ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_n(\boldsymbol{X})) - \ell(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\theta}_0(\boldsymbol{X}))\right].
\end{aligned}$$

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

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

We now apply the localization-based analysis of Farrell et al. (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 et al. (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 (A.1) is bounded as

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

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 et al. (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)

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

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 et al. (2021) to the term for each component function θ_k, $k = 1,\dots,d_{\boldsymbol{\theta}}$, yielding

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

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

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

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

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

$$\begin{aligned}&c_1\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0(\boldsymbol{X})\|_2^2\right] \\ &\quad\le 6Kr_0\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{DNN}})}{n}\log n} + \sqrt{\frac{2C_\ell^2 r_0^2\gamma}{n}} + \frac{23\cdot 3MC_\ell}{3}\frac{\gamma}{n} + c_2\epsilon_n^2 + \epsilon_n\sqrt{\frac{2C_\ell^2\gamma}{n}} + \frac{7C_\ell M\gamma}{n} \\ &\quad\le r_0\left(6K\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{DNN}})}{n}\log n} + \sqrt{\frac{2C_\ell^2\gamma}{n}}\right) + c_2\epsilon_n^2 + \epsilon_n\sqrt{\frac{2C_\ell^2\gamma}{n}} + K_2\frac{\gamma}{n} \\ &\quad\le r_0\left(K_1\sqrt{\frac{WL\log(W)}{n}\log n} + \sqrt{\frac{2C_\ell^2\gamma}{n}}\right) + c_2\epsilon_n^2 + \epsilon_n\sqrt{\frac{2C_\ell^2\gamma}{n}} + K_2\frac{\gamma}{n},\end{aligned} \tag{A.3}$$

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 (A.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 et al. (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 $1 - e^{-\gamma_1}$,

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

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 et al. (2021). This result tell us that for each θ_{0k}, the following holds for H, L, and the approximation error ϵ_n:

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

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 $x_{d_{\boldsymbol{X}}}$ is binary. Then for two smooth, $d_{\boldsymbol{X}} - 1$-dimensional functions $\theta_{0k,1}$ and $\theta_{0k,0}$, it holds that $\theta_{0k}(\boldsymbol{x}) = x_{d_{\boldsymbol{X}}}\theta_{0k,1}(x_1,\ldots,x_{d_{\boldsymbol{X}}-1}) + (1-x_{d_{\boldsymbol{X}}})\theta_{0k,0}(x_1,\ldots,x_{d_{\boldsymbol{X}}-1})$. Adding a single node to each hidden layer allows the network to pass forward the input $x_{d_{\boldsymbol{X}}}$ and multiply it with two separate learned functions just prior to the parameter, giving exactly $\bar{\theta}_{k,n}(\boldsymbol{x}) = x_{d_{\boldsymbol{X}}}\bar{\theta}_{0k,1}(x_1,\ldots,x_{d_{\boldsymbol{X}}-1}) + (1-x_{d_{\boldsymbol{X}}})\bar{\theta}_{0k,0}(x_1,\ldots,x_{d_{\boldsymbol{X}}-1})$. Intuitively, this is like Figure 2, with $x_{d_{\boldsymbol{X}}}$ in place of $\boldsymbol{t}$ and the two functions $\theta_{0k,1}$ and $\theta_{0k,0}$ in the parameter layer (and then feeding into the appropriate output). The same argument can be applied to every category of the discrete data and to each function to be learned. The estimator matches our structure exactly. Since $d_{\boldsymbol{X}}$ is fixed, this results in only a constant increase in the width of the network. Put together, we take $\epsilon_n = n^{-\frac{\beta}{2(\beta+d)}}$, i.e. $H \asymp n^{\frac{d}{2(\beta+d)}}\log^2 n$, $L \asymp \log n$, and we obtain the final result. □

A.2 Proof of Corollary 1

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 $G^{-1}\left(\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}]\right) = \boldsymbol{\theta}_1(\boldsymbol{x})'\boldsymbol{t} = \boldsymbol{\theta}_2(\boldsymbol{x})'\boldsymbol{t}$ a.e. or equivalently that for $\boldsymbol{\theta}_*(\boldsymbol{x}) = \boldsymbol{\theta}_1(\boldsymbol{x}) - \boldsymbol{\theta}_2(\boldsymbol{x})$, $\boldsymbol{\theta}_*(\boldsymbol{x})'\boldsymbol{t} = 0$. But $\boldsymbol{\theta}_*(\boldsymbol{x})'\boldsymbol{t} = 0$ a.e. implies that

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

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 3. The fact that $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}' \mid \boldsymbol{X}]$ is (uniformly) positive yields

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

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 $g(G(\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t}))$. 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 $g(G(\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t}))$, and the smoothness of these functions mean that the linear index $\boldsymbol{\theta}(\boldsymbol{x})'\boldsymbol{t}$ can be recovered. □

B Proof of Theorem 2

Here we derive the influence function for $\boldsymbol{\mu}_0$. Recall that our purpose is to derive an influence function to use as a basis for estimation and inference, in particular to obtain a Neyman orthogonal score, not in efficiency characterizations or other theory. For in-depth treatments, including discussion of regularity conditions, efficiency bounds, and other concerns, see Newey (1990), Newey (1994), van der Vaart (1998, Chapter 25) and Ichimura and Newey (2015). In particular, we apply the pathwise derivative approach as detailed by Newey (1994).

The starting point is a parametric submodel, indexed by a parameter η. Distributions and other nonparametric objects are indexed by η, and thus we define $\boldsymbol{\theta}(\boldsymbol{x};\eta)$ and $\boldsymbol{\mu}_0(\eta)$ as

$$\boldsymbol{\theta}(\cdot;\eta) = \arg\min_{\boldsymbol{b}} \int \ell\left(\boldsymbol{w}, \boldsymbol{b}(\boldsymbol{x})\right) f_{\boldsymbol{w}}(\boldsymbol{w};\eta) d\boldsymbol{w} \tag{B.1}$$

and

$$\boldsymbol{\mu}(\eta) = \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};\eta); \boldsymbol{t}^*\right) f_{\boldsymbol{x}}(\boldsymbol{x};\eta) d\boldsymbol{x}, \tag{B.2}$$

where $f_{\boldsymbol{w}}$ and $f_{\boldsymbol{x}}$ are the distributions of $\boldsymbol{w} = (\boldsymbol{y}', \boldsymbol{t}', \boldsymbol{x}')'$ and $\boldsymbol{x}$ respectively. 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

$$\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{W})S(\boldsymbol{W})], \tag{B.3}$$

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

The first step is differentiating (B.2) with respect to the parameter η, and evaluating this at

$\eta = 0$. The product rule and the chain rule yield

$$\begin{aligned}
\left.\frac{\partial \boldsymbol{\mu}(\eta)}{\partial \eta}\right|_{\eta=0} &= \left.\frac{\partial}{\partial \eta}\left\{\int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};\eta); \boldsymbol{t}^*\right) f_{\boldsymbol{x}}(\boldsymbol{x};\eta) d\boldsymbol{x}\right\}\right|_{\eta=0} \\
&= \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};0); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} + \int \left.\frac{\partial \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x};\eta); \boldsymbol{t}^*\right)}{\partial \eta}\right|_{\eta=0} f_{\boldsymbol{x}}(\boldsymbol{x};0) d\boldsymbol{x}, \\
&= \int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} + \int \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*) \boldsymbol{\theta}_\eta(\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}, \qquad \text{(B.4)}
\end{aligned}$$

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

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

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 (B.4) above can be written as expectations of products with the full score $S(y, \boldsymbol{x}, \boldsymbol{t})$, as required by (B.3). We will often use the standard facts that scores are mean zero and that

$$S(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{t}) = S(\boldsymbol{y}, \boldsymbol{t} \mid \boldsymbol{x}) + S(\boldsymbol{x}). \qquad \text{(B.5)}$$

The first term of Equation (B.4) is

$$\begin{aligned}
\int \boldsymbol{H}\left(\boldsymbol{x}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) \left.\frac{\partial f_{\boldsymbol{x}}(\boldsymbol{x};\eta)}{\partial \eta}\right|_{\eta=0} d\boldsymbol{x} &= \mathbb{E}\left[\boldsymbol{H}\left(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) S(\boldsymbol{X})\right] \\
&= \mathbb{E}\left[\boldsymbol{H}\left(\boldsymbol{X}, \boldsymbol{\theta}_0(\boldsymbol{x}); \boldsymbol{t}^*\right) S(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{T})\right], \qquad \text{(B.6)}
\end{aligned}$$

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 (B.5):

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

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 $\widehat{\boldsymbol{\beta}}(\boldsymbol{x})$ were fixed). The second term of Equation (B.4) 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 (B.4) 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,

$$\mathbb{E}_\eta\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W},\boldsymbol{\theta}(\boldsymbol{x};\eta))\middle|\,\boldsymbol{X}=\boldsymbol{x}\right]\equiv 0, \tag{B.7}$$

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

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

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

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

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 $\ell(\boldsymbol{w},\boldsymbol{b})$ 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

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

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

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

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

$$\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y},\boldsymbol{T}\mid\boldsymbol{X})\middle|\,\boldsymbol{X}\right]+\mathbb{E}\left[\boldsymbol{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}_0(\boldsymbol{x}))\boldsymbol{\theta}_\eta(\boldsymbol{x})\middle|\,\boldsymbol{X}\right]=0, \tag{B.8}$$

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 (B.8), 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 (B.4) 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 (B.5), the second term of Equation (B.4) is of the required form:

$$-\mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{X});\boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W},\boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{X})\Big] \tag{B.9}$$

Combining Equations (B.6) and (B.9) with (B.4), we find that

$$\begin{aligned}\frac{\partial\boldsymbol{\mu}(\eta)}{\partial\eta}\bigg|_{\eta=0} &= \mathbb{E}\left[\boldsymbol{H}\left(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*\right)S(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{T})\right]\\ &\quad-\mathbb{E}\Big[\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{\theta}_0(\boldsymbol{X});\boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{X})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{W},\boldsymbol{\theta}_0(\boldsymbol{x}))S(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{X})\Big].\end{aligned} \tag{B.10}$$

Thus we have verified Equation (B.3) with

$$\boldsymbol{\psi}(\boldsymbol{w}) = \boldsymbol{H}\left(\boldsymbol{x},\boldsymbol{\theta}_0(\boldsymbol{x});\boldsymbol{t}^*\right) - \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}_0(\boldsymbol{X});\boldsymbol{t}^*)\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\boldsymbol{\ell}_{\boldsymbol{\theta}}(\boldsymbol{w},\boldsymbol{\theta}_0(\boldsymbol{x})). \tag{B.11}$$

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 (B.3).

C Proof of Theorem 3: Asymptotic Normality

The result follows from Theorems 3.1 and 3.2 of Chernozhukov et al. (2018) upon verifying Assumptions 3.1 and 3.2 therein. Assumption 3.1(a) holds by for $\boldsymbol{\psi} - \boldsymbol{\mu}_0$ given in Theorem 2: the first term of $\boldsymbol{\psi}$ has mean $\boldsymbol{\mu}_0$ by (2.2) while the second is (conditionally) mean zero as assumed in Assumption 4, with $\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}$ uniformly bounded. Assumption 3.1(b), linearity, holds by definition of (2.2) and the form of the score in Theorem 2. Assumption 3.1(c) holds by Assumption 4, 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 Theorem 2. 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. (2018) and the assumed convergence of the first stage estimates of Assumption 5.