Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands[*]

Max H. Farrell Tengyuan Liang Sanjog Misra

University of Chicago, Booth School of Business

October 2, 2018

Abstract

We study deep neural networks and their use in semiparametric inference. We provide new rates of convergence for deep feedforward neural nets and, because our rates are sufficiently fast (in some cases minimax optimal), prove that semiparametric inference is valid using deep nets for first-step estimation. Our estimation rates and semiparametric inference results are the first in the literature to handle the current standard architecture: fully connected feedforward neural networks (multi-layer perceptrons), with the now-default rectified linear unit (ReLU) activation function and a depth explicitly diverging with the sample size. We discuss other architectures as well, including fixed-width, very deep networks. We establish nonasymptotic bounds for these deep ReLU nets, for both least squares and logistic loses in nonparametric regression. We then apply our theory to develop semiparametric inference, focusing on treatment effects and expected profits for concreteness, and demonstrate their effectiveness with an empirical application to direct mail marketing. Inference in many other semiparametric contexts can be readily obtained.

Keywords: Deep Neural Networks, Rectified Linear Unit, Nonasymptotic Bounds, Semiparametric Inference, Treatment Effects, Program Evaluation, Treatment Targeting.

1 Introduction

Statistical machine learning methods are being rapidly integrated into the social and medical sciences. These methods have found a variety of applications, but have been particularly useful in "big data" settings and studying casual effects. Furthermore, theoretical properties of these methods are the subject of intense recent study. This has netted several breakthroughs both theoretically, such as robust, valid inference following machine learning, and in novel applications and conclusions. Our goal in the present work is to study a particular statistical machine learning technique which is widely popular in industrial applications, but infrequently used in academic work and largely

[*] We thank Milica Popovic for outstanding research assistance. Liang gratefully acknowledges support from the George C. Tiao Fellowship. Misra gratefully acknowledges support from the Neubauer Family Foundation.

ignored in recent theoretical developments on inference: deep neural networks. Neural networks are perhaps not as familiar to social scientists, and indeed, were even out of favor in the machine learning community for several years, returning to prominence only very recently. Our work aims to bring wider attention to these methods and to take the first step toward filling the gap in theoretical understanding of inference using deep neural networks. We derive rates of convergence for deep nets and show that these can be used to obtain valid inference for semiparametric objects, focusing on causal effects in particular. To our knowledge this is the first inference result using deep nets.

Before the recent surge in attention, neural networks had taken a back seat to other methods (such as kernel methods or forests) largely because of their modest empirical performance and challenging optimization. Recently however, *deep* neural networks, with many hidden layers, have been found to perform extremely well, matching or setting the state of the art (Krizhevsky et al., 2012; He et al., 2016). Optimization problems have seemingly been overcome by modern computational power, stochastic optimization techniques (LeCun et al., 1998; Kingma and Ba, 2014), and the change from smooth sigmoid-type activation functions to rectified linear units (ReLU), $x \mapsto \max(x, 0)$ (Nair and Hinton, 2010). Our results speak directly to this modern implementation. We explicitly model the depth of the network as diverging with the sample size and focus on the ReLU activation function.

Further back in history, before falling out of favor, neural networks were widely studied and applied, particularly around the 1990s. In that time, *shallow* neural networks were shown to have many good theoretical properties. Intuitively, neural networks are a form of sieve estimation, wherein basis functions of the original variables are used to approximate unknown nonparametric objects. What sets neural nets apart is that the basis functions are themselves learned from the data by optimizing over many flexible combinations of simple functions. It has been known for some time that such networks yield universal approximations (Hornik et al., 1989). Comprehensive theoretical treatments are given by White (1992) and Anthony and Bartlett (1999). Of particular relevance in this strand of theoretical work are Chen and Shen (1998) and Chen and White (1999), as well as references therein, where it was shown that single-layer, sigmoid-based networks could attain sufficiently fast rates for semiparametric inference (see Chen (2007) for more references in econometrics).

We explicitly depart from these works by focusing on the modern setting of deep neural networks with the rectified linear (ReLU) activation function. We briefly review their construction in Section 2; for more see Goodfellow et al. (2016). Our main theoretical contributions are nonasymptotic bounds for nonparametric estimation using deep ReLU networks, for both least squares and logistic regression. These bounds immediately imply convergence rates. The bounds and convergence rates appear to be new to the literature. We derive these using a novel localization analysis, which delivers sharp bounds and hence fast rates. Recent developments in approximation theory (Yarotsky, 2017, 2018) and complexity (Bartlett et al., 2017) for deep ReLU networks are important building blocks for our results.

Our second main result is valid inference on semiparametric causal effects following first-step estimation using deep nets. Program evaluation with observational data is one of the most common and important semiparametric inference problems, and has often been used as a first test case for theoretical study of inference following machine learning (e.g., Belloni et al., 2014; Farrell, 2015; Belloni et al., 2017; Athey et al., 2016). (Causal inference as a whole is a vast literature; see Imbens and Rubin (2015) for a broad review and Abadie and Cattaneo (2018) for a recent review of program evaluation methods, and further references in both.) Our results cover many interesting causal estimands, and we focus in particular on average treatment effects and expected profits from treatment targeting strategies. Our results allow planners (e.g., firms or medical providers) firms to compare different strategies, either predetermined or estimated using auxiliary data, and recognizing that targeting can be costly, decide which strategy to implement. Other estimands are discussed briefly. Deep neural networks have been argued (experimentally) to outperform the previous state-of-the-art in causal inference (Westreich et al., 2010; Johansson et al., 2016; Shalit et al., 2017; Hartford et al., 2017). To the best of our knowledge, ours are among the first theoretical results that explicitly deliver inference using deep neural networks.

We focus on causal effect type parameters for concreteness and their wide applicability, as well as to allow direct comparison to the literature above, but our results are not limited to only these objects. In particular, our results yield inference on essentially any estimand that admits a locally robust estimator (Chernozhukov et al., 2018b) depending only on conditional expectations (under appropriate regularity conditions). Our aim is not to innovate at the semiparametric step, for example by seeking weaker conditions on the first stage, but rather, we aim to utilize such results. Our work contributes directly to this area of research by showing that deep nets are a valid and useful first-step estimator, in particular, attaining a rate of $o(n^{-1/4})$ under appropriate smoothness conditions.

We illustrate our results, and more generally the utility of deep ReLU networks, by studying causal effects in the context of a direct mail marketing campaign. The data come from a large US consumer products retailer and consists of close to three hundred thousand consumers with one hundred fifty covariates. This is the same data that was used by Hitsch and Misra (2018) to study various estimators, both traditional and modern, of heterogeneous treatment effects. We refer the reader to that paper for a more complete description of the data as well as results using other estimators (see also Hansen et al. (2017)). We study the effect of catalog mailings on consumer purchases, and moreover, compare different targeting strategies (i.e. to which consumers catalogs should mailed). The cost of sending out a single catalog can be close to one dollar, and with millions being set out, carefully assessing the targeting strategy is crucial. Our results suggest that deep nets are at least as good as (and sometimes better) that the best methods found by Hitsch and Misra (2018).

The remainder of the paper proceeds as follows. Next, we briefly review the related theoretical literature. Section 2 introduces deep ReLU networks and states our main theoretical results: nonasymptotic bounds and convergence rates for least squares and logistic regression. Section 3 gives details for semiparametric causal inference. The empirical application is presented in Section 4. Section 5 concludes. All proofs are given in the appendix.

1.1 Related Theoretical Literature

Our paper contributes to several rapidly growing literatures, and we can not hope to do justice to each here. We give only those of particular relevance; more references can be found in these works. First, there has been much recent theoretical study of the properties of the machine learning tools, either as an end in itself or with an eye toward use in semiparametric inference. Much of this work has focused on the lasso and its variants (Bickel et al., 2009; Belloni et al., 2011, 2012;

Farrell, 2015) and tree/forest based methods (Wager and Athey, 2018), though earlier work studied shallow (typically with a single hidden layer) neural networks with smooth activation functions (White, 1989, 1992; Chen and White, 1999). We fill the gap in this literature by studying *deep* neural networks with ReLU activation, which is non-smooth.

A second, intertwined strand of literature focuses on inference following the use of machine learning methods, often with a focus on causal effects. Initial theoretical results were concerned obtaining valid inference on a coefficient in a high-dimensional regression, following model selection or regularization, with particular focus on the lasso (Belloni et al., 2012; Javanmard and Montanari, 2014; van de Geer et al., 2014). Intuitively, this is a semiparametric problem, where the coefficient of interest is estimable at the parametric rate, and the remaining coefficients are a nonparametric nuisance parameter estimated using machine learning methods. Building on this intuition, many have studied the semiparametric stage directly, such as obtaining novel, weaker conditions easing the application of machine learning methods (Belloni et al., 2014; Farrell, 2015; Chernozhukov et al., 2018b; Belloni et al., 2018, and references therein). Conceptually related to this strand are targeted maximum likelihood (van der Laan and Rose, 2001) and the higher-order influence functions (Robins et al., 2008, 2017). Our work builds on this work, employing conditions therein, and in particular, verifying them for deep ReLU nets.

Finally, our convergence rates build on, and contribute to, the recent theoretical machine learning literature on deep neural networks. Because of the renaissance in deep learning, a considerable amount of study has been done in recent years. Of particular relevance to us are Yarotsky (2017, 2018) and Bartlett et al. (2017); a recent textbook treatment, containing numerous other references is given by Goodfellow et al. (2016).

2 Deep ReLU Networks

In this section we will give our main theoretical results: nonasymptotic bounds and associated convergence rates for nonparametric least squares and logistic regression using deep ReLU nets. The utility of these results for second-step semiparametric causal inference (the downstream task), for which our rates are sufficiently rapid, is demonstrated in Section 3. We view our results as an initial step in establishing both the estimation and inference theory for modern deep neural networks, i.e. those built using a multi-layer perceptron architecture (described below) and the nonsmooth ReLU activation function, the combination of which has demonstrated state of the art performance empirically and can be feasibly optimized. As a note on exposition, while our main results are in fact nonasymptotic bounds that hold with high probability, for simplicity we will refer our results as "rates" in most discussion.

As neural networks are perhaps less familiar to applied social scientists, we first briefly review the construction of deep ReLU nets. Our main focus will be on the standard fully connected feedfoward neural network, frequently referred to as a multi-layer perceptron, as this is the most commonly implemented network architecture and we want our results to inform empirical practice. However, our results are more general, accommodating other architectures provided they are able to yield a universal approximation (in the appropriate function class), and so we review neural nets more generally and give concrete examples.

The nonparametric setup we consider is standard. Our goal is to estimate an unknown smooth function $f_*(\boldsymbol{x})$, that relates the covariates $\boldsymbol{X} \in \mathbb{R}^d$ to a scalar outcome Y. We collect these random variables into the vector $\boldsymbol{Z} = (Y, \boldsymbol{X}')' \in \mathbb{R}^{d+1}$ and let a realization be $\boldsymbol{z} = (y, \boldsymbol{x}')'$ Accuracy is measured by a per-observation loss function denoted by $\ell(f, \boldsymbol{z})$. We study least squares and logistic regression. Both play a role in semiparametric inference on causal effects, in particular corresponding to the outcome and propensity score models, respectively. For least squares, the target function and loss are

$$f_*(\boldsymbol{x}) := \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] \qquad \text{and} \qquad \ell\left(f, \boldsymbol{z}\right) = \frac{1}{2}(y - f(\boldsymbol{x}))^2, \tag{2.1}$$

respectively, while for logistic regression these are

$$f_*(\boldsymbol{x}) := \log \frac{\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]}{1 - \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]} \qquad \text{and} \qquad \ell\left(f, \boldsymbol{z}\right) = -yf(\boldsymbol{x}) + \log\left(1 + e^{f(\boldsymbol{x})}\right). \tag{2.2}$$

Other loss functions can be accommodated; our results naturally extend to cases such as generalized linear models, for example multinomial logistic regression.

2.1 Neural Network Constructions

For either least squares or logistic loss, we estimate the target function using a deep ReLU network. We will give a brief outline of their construction here, paying closer attention to the details germane to our theory; a more complete introduction is given by Goodfellow et al. (2016).

The crucial choice is the specific network architecture, or class. In general we will call this $\mathcal{F}_{\text{DNN}}$. From a theoretical point of view, different classes have different complexity and different approximating power. We give results for several concrete examples below. We will focus on *feedforward neural networks* (Anthony and Bartlett, 1999). An example of a feedforward network is shown in Figure 1. The network consists of d input units, that correspond to covariates $\boldsymbol{X} \in \mathbb{R}^d$, one output unit for the outcome Y, and between them, U hidden units or computational nodes. These are connected by a directed acyclic graph specifying the architecture, and the description is completed by the choice of an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. In this paper, we focus on the popular ReLU activation function $\sigma(x) = \max(x, 0)$.

The key structural feature of a feedforward network is that hidden units are grouped in a sequence of L layers, the *depth* of the network, where a node is in layer $l = 1, 2, \ldots, L$, if it has a predecessor in layer $l - 1$ and no predecessor in any layer $l' \geq l$. Computation of the final output unit proceeds layer-by-layer: at any layer $l \leq L$, each hidden unit u receives an input in the form of a linear combination $\tilde{\boldsymbol{x}}'\boldsymbol{w} + b$, and then returns (outputs) $\sigma(\tilde{\boldsymbol{x}}'\boldsymbol{w} + b)$, where the vector $\tilde{\boldsymbol{x}}$ collects the output of all the hidden units with a directed edge into u (i.e., from prior layers). The dimension of $\tilde{\boldsymbol{x}}$ is the *width* of the network at that point, and is not equal to d, the dimension of $\boldsymbol{X}$, but rather varies according the network structure leading into unit u. Finally, for the output layer the activation function is the identity, and thus the final output is $\tilde{\boldsymbol{x}}'\boldsymbol{w} + b$.

The parameters are the weight vector $\boldsymbol{w}$ and the constant term b, with one set $(\boldsymbol{w}, b)$ for each node of the graph. (The constant term is often referred to as the "bias" in the network literature, but given the loaded meaning of this term in inference, we will avoid referring to b as a bias.) The collection, over all nodes, of $\boldsymbol{w}$ and b, constitutes the parameters θ which are optimized in the final estimation. We denote W as the total number of parameters (both weights and biases) of the network. In practice, this optimization problem is solved efficiently using variants of stochastic gradient descent

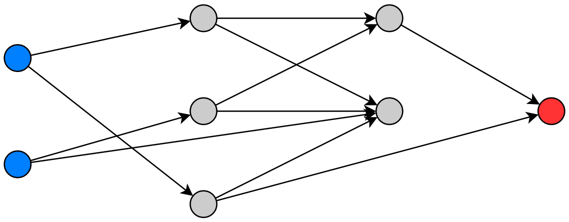


Figure 1: Illustration of a feedforward neural network with $W = 17$, $L = 2$, $U = 5$, and input dimension $d = 2$. Blue ones are the input units, grey for hidden units, and red is the output unit.

algorithm, with stochastic gradients of the parameters calculated by back-propagation (chain-rule) induced by the network structure.

In sum, for a user-chosen architecture $\mathcal{F}_{\text{DNN}}$, encompassing the choices $\sigma(\cdot)$, U, L, W, and the graph structure, the final estimate is computed using observed samples $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i')'$, $i = 1, 2, \ldots, n$, of $\boldsymbol{Z}$, by solving, for an $M > 0$,

$$\widehat{f}_{\text{DNN}} := \underset{\substack{f_\theta \in \mathcal{F}_{\text{DNN}} \\ \|f_\theta\|_\infty \leq 2M}}{\arg\min} \sum_{i=1}^{n} \ell\left(f, \boldsymbol{z}_i\right). \tag{2.3}$$

An important and widely used subclass is the one that is *fully connected* between consecutive layers but has *no* other connections and each layer has number of hidden units that are of the same order of magnitude. This architecture is often referred to as a *Multi-Layer Perceptron* (MLP). See Figure 2, cf. Figure 1. For each of the L hidden layers, let H be the number of hidden units, the *width* of the network. For the present exposition, assume all layers have the same width H; below $H = H_n$ will be the growth rate of the width, assumed common across layers. We denote the class as $\mathcal{F}_{\text{MLP}}$. For this class, $U = LH$ and $W = (d+1)H + (L-1)(H^2 + H) + H + 1$. When (2.3) is restricted to this class we denote the resulting estimator $\widehat{f}_{\text{MLP}}$. We will allow for generic feedforward networks in our results, but we present special results for the MLP case, as it is widely used in empirical practice. As we will see below, the architecture, through its complexity, and more importantly, approximation power, plays a crucial role in the final convergence rate. In particular, the rate obtained for the MLP case will be suboptimal, though still sufficiently rapid for semiparametric inference.

To further clarify the class of deep nets, it is useful to make explicit analogies to more classical

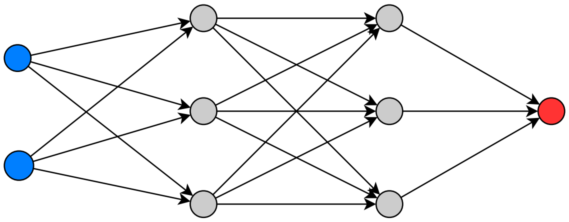


Figure 2: Illustration of multi-layer perceptron $\mathcal{F}_{\text{MLP}}$ with $H = 3$, $L = 2$ ($U = 6$, $W = 25$), and input dimension $d = 2$.

nonparametric techniques.. For exposition, suppose $Y \in \mathbb{R}$ is continuous (thus we focus on least squares (2.1)) and we use a multi-layer perceptron. Let the output of the computational node $u = (h, l)$, $h = 1, \dots H$, $l = 1, \dots L$, be the scalar $\tilde{x}_{h,l}$; this is given by $\tilde{x}_{h,l} = \sigma(\tilde{\boldsymbol{x}}'_{l-1}\boldsymbol{w}_{h,l-1} + b_{h,l-1})$, where $(\tilde{\boldsymbol{x}}_{l-1} = (\tilde{x}_{1,l-1}, \dots, \tilde{x}_{H,l-1})'$. Then the final output is a linear combination of these: $\hat{y} = \tilde{\boldsymbol{x}}'_L\boldsymbol{w}_L + b_L$. Once we recall that $\tilde{\boldsymbol{x}}_L = \tilde{\boldsymbol{x}}_L(\boldsymbol{x})$, we can view this as a basis function expansion (albeit a complex one) of the original observation $\boldsymbol{x}$, and then the form $\hat{f}_{\text{MLP}}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}_L(\boldsymbol{x})'\boldsymbol{w}_L + b_L$ is reminiscent of a traditional series (linear sieve) estimator. The crucial distinction now is that the set of basis functions $\tilde{\boldsymbol{x}}_L(\cdot)$ is learned from the data.

For a traditional series estimator, say smoothing splines, the two choices for the practitioner are the spline basis (the shape and the degree) and the number of terms (knots), commonly referred to as the smoothing and tuning parameters, respectively. In kernel regression, these would respectively be the shape of the kernel (and degree of local polynomial) and the bandwidth(s). For neural networks, the same phenomena are present: the architecture as a *whole* (the graph structure and activation function) are the smoothing parameters while the width and depth play the role of tuning parameters.

The architecture plays a crucial role in that it determines the approximation power of the network, and it is worth noting that because of the relative complexity of neural networks, such approximations, and comparisons across architectures, are not as simple. It is comparatively obvious that quartic splines are more flexible than cubic splines (for the same number of knots) as is a higher degree local polynomial (for the same bandwidth). At a glance, it may not be clear what a given network architecture (width, depth, graph structure, and activation function) can approximate. As we will show below, the MLP architecture is not yet known to yield an optimal approximation (for a given width and depth) and therefore attains a slower rate. As a final note, computational considerations are important for deep nets in a way that is not true conventionally. The activation function is crucial in this regard. The recent switch to ReLU stems not from their greater approximation power, but from the fact that optimizing a deep net with sigmoid-type activation is unstable or impossible in practice. It is for these reasons that we focus on ReLU networks and also that we present both generic results and specific corollaries for different architectures.

Just as for classical nonparametrics, for a fixed architecture, it is the tuning parameter choices that determine the rate of convergence (for a fixed smoothness of the underlying function). The new wave of study of neural networks, focusing on depth, is in its infancy theoretically. As such, there is no understanding yet of optimal architecture(s) or tuning parameters. Choices of both are quite difficult, and only preliminary research has been done (e.g., Daniely, 2017; Telgarsky, 2016; Safran and Shamir, 2016; Mhaskar and Poggio, 2016a; Raghu et al., 2017, and references therein). Further exploration of these ideas is beyond the current scope. It is interesting to note that in some cases, a good approximation can be obtained even with a fixed width H, provided the network is deep enough, a very particular way of enriching the "sieve space" $\mathcal{F}_{\text{DNN}}$. See Corollary 2.

Remark 1. In applications it is common to apply some form of regularization to the optimization of (2.3). However, in theory, the role of explicit regularization is unclear and may be unnecessary, as stochastic gradient descent presents good, if not better, solutions empirically (Zhang et al., 2016). Regularization may improve empirical performance in low signal-to-noise ratio problems. A detailed investigation is beyond the scope of the current work, though we do investigate this numerically in Section 4. There are many alternative regularization methods, including L_1 and L_2 (weight decay) penalties, drop out, and others. ⌟

2.2 Bounds and Convergence Rates for Multi-Layer Perceptrons

We can now state our main theoretical results: bounds and convergence rates for deep ReLU networks. All proofs appear in Appendix A. We study neural networks from a nonparametric point of view (e.g., White, 1989, 1992; Schmidt-Hieber, 2017; Liang, 2017; Bauer and Kohler, 2017, in specific scenarios). Chen and Shen (1998) and Chen and White (1999) share our goal, fast convergence rates for use in semiparametric inference, but focus on shallow, sigmoid-based neural networks compared to our deep ReLU nets, though also allowing for dependent data. They are able to attain fast rates by appealing to stronger approximation results by Barron (1993), whereas for our analysis of deep nets we employ localization (Koltchinskii and Panchenko, 2000; Bartlett et al., 2005; Koltchinskii, 2006; Liang et al., 2015). We use recent approximation work of Yarotsky (2017, 2018) and complexity results by Bartlett et al. (2017), together with this localization.

The regularity conditions we require are collected in the following.

Assumption 1. *Assume that $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i')'$, $1 \leq i \leq n$ are i.i.d. copies of $\boldsymbol{Z} = (Y, \boldsymbol{X}) \in \mathcal{Y} \times [-1,1]^d$, where X is continuously distributed. For an absolute constant $M > 0$, assume $\|f_*\|_\infty \leq M$ and either $f_*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ and $\mathcal{Y} = [-M, M]$ or $f_*(\boldsymbol{x}) = \log\left(\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]/(1 - \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}])\right)$ and $\mathcal{Y} = \{0, 1\}$.*

This assumption is fairly standard in nonparametrics. The only restriction worth mentioning is that for least squares regression we assume that the outcome is bounded. This is a stronger condition that is often found, and we conjecture that it can be relaxed by assuming $Y = f_*(\boldsymbol{X}) + \varepsilon$, for a stochastic error ε that has sufficiently many moments or other distributional requirements. In common practice our restriction is not substantially more limiting than such an assumption when coupled with the standard boundedness of f_*. The assumption of continuously distributed covariates is quite standard. From a theoretical point of view, covariates taking on only a few values can be conditioned on and then averaged over, and these will, as usual, not enter into the dimensionality which curses the rates. Discrete covariates taking on many values may be more realistically thought of as continuous, and it may be more accurate to allow these to slow the convergence rates. Our focus on $L_2(X)$ convergence allows for these essentially automatically. Finally, from a practical point of view, deep networks handle discrete covariates seamlessly and have demonstrated excellent empirical performance, which is in contrast to other more classical nonparametric techniques that may require more manual adaptation.

We begin with the most important network architecture, the multi-layer perceptron case. This is the most widely used network architecture in practice and an important contribution of our work is to cover this directly, along with ReLU activation. MLPs are known now to approximate smooth functions well, leading to our next assumption: that the target function f_* lies in a Sobolev ball with certain smoothness.

Assumption 2. *Assume f_* lies in the Sobolev ball $\mathcal{W}^{\beta,\infty}([-1,1]^d)$, with smoothness $\beta \in \mathbb{N}_+$,*

$$f_*(x) \in \mathcal{W}^{\beta,\infty}([-1,1]^d) := \left\{ f : \max_{\alpha, |\alpha| \leq \beta} \operatorname*{ess\,sup}_{x \in [-1,1]^d} |D^\alpha f(x)| \leq 1 \right\},$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha| = \alpha_1 + \ldots + \alpha_d$ and $D^\alpha f$ is the weak derivative.

Under Assumptions 1 and 2 we obtain the following result, which, to the best of our knowledge, is new to the literature. In some sense, this is our main result for deep ReLU networks, as it deals with the most common architecture. We apply this in Sections 3 and 4 below for causal inference.

Theorem 1 (Multi-Layer Perceptron)**.** *Suppose Assumptions 1 and 2 hold. Let $\widehat{f}_{\mathrm{MLP}}$ be the deep ReLU network estimator defined by (2.3), restricted to $\mathcal{F}_{\mathrm{MLP}}$, for either least squares (2.1) or logistic (2.2) loss, with $H_n \asymp n^{\frac{d}{2(\beta+d)}} \log^2 n$ and $L_n \asymp \log n$. Then with probability at least $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$,*

(a) $\|\widehat{f}_{\mathrm{MLP}} - f_*\|^2_{L_2(X)} \leq C \cdot \left\{ n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log \log n}{n} \right\}$ *and*

(b) $\mathbb{E}_n \left[(\widehat{f}_{\mathrm{MLP}} - f_*)^2 \right] \leq C \cdot \left\{ n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log \log n}{n} \right\}$,

for a universal constant $C > 0$ independent of n.

The proof of this result is deferred to the Appendix. However, several aspects warrant discussion. We build on the recent results of Bartlett et al. (2017), who find nearly-tight bounds on the Vapnik-Chervonenkis (VC) dimension of deep nets. One contribution of our proof is to derive a *scale sensitive* localization theory with *scale insensitive* measures, such as VC- or Pseudo-dimension, for deep neural networks. This has two tangible benefits. First, we do not restrict the class of network architectures to have bounded weights for each unit (scale insensitive), in accordance to standard practice (Zhang et al., 2016). Moreover, this allows for a richer set of approximating possibilities, in particular allowing more flexibility in seeking architectures with specific properties, as we explore in the next subsection. This is in contrast to the classic sieve analysis with scale sensitive measure such as metric entropy (used by Chen and White, 1999, for example). Second, from a technical point of view, we are able to attain a faster rate on the second term of the bound, order n^{-1} in the sample size, instead of the $n^{-1/2}$ that would result from a direct application of uniform deviation bounds. This upper bound informs the trade offs between H_n and L_n, and the approximation power, and may point toward optimal architectures for statistical inference.

This result gives a nonasymptotic bound that holds with high probability. As mentioned above, we will generally refer to our results simply as "rates" when this causes no confusion. This result relies on choosing H_n appropriately given the smoothness β of Assumption 2. Of course, the true smoothness is unknown and thus in practice the "β" appearing in H_n, and consequently in the convergence rates, need not match that of Assumption 2. In general, the rate will depend on the smaller of the two. Most commonly it is assumed that the user-chosen β is fixed and that the truth is much smoother; witness the ubiquity of cubic splines and local linear regression. Rather than spell out these consequences directly, we will tacitly assume the true smoothness is not less than the β appearing in H_n (here and below). Adaptive approaches, as in classical nonparametrics, may also be possible with deep nets, but are beyond the scope of this study.

Even with these choices of H_n and L_n, the rate of Theorem 1 is not optimal (for fixed β, in the sense of Stone (1982)). We rely on the explicit approximating constructions of Yarotsky (2017), and it is possible that in the future improved approximation properties of MLPs will be found, allowing for a sharpening of Theorem 1 without substantive change to the argument.

2.3 Other Network Architectures

Theorem 1 covers only one specific architecture, albeit the most import one. However, given that this field is rapidly evolving, it is important to consider other possible architectures which may be beneficial in some cases. To this end, we will state a more generic result and then two specific examples: one to obtain an faster rate of convergence and one for fixed-width networks. All of these results are, at present, more of theoretical interest than practical value, as they are either agnostic about the network (thus infeasible) or rely on more limiting assumptions.

In order to be agnostic about the specific architecture of the network we need to be flexible in the approximation power of the class. To this end, we will replace Assumption 2 with the following generic assumption, rather more of a definition, regarding the approximation power of the network.

Assumption 3. *Let f_* lie in a class $\mathcal{F}$. For the feedforward network class $\mathcal{F}_{\mathrm{DNN}}$, used in* (2.3)*,* *the approximation error* ϵ_{DNN} *is*

$$\epsilon_{\text{DNN}} := \sup_{f_* \in \mathcal{F}} \inf_{\substack{f \in \mathcal{F}_{\text{DNN}} \\ \|f\|_\infty \leq 2M}} \|f - f_*\|_\infty \ .$$

Under this condition we obtain our the following generic result.

Theorem 2 (General Feedforward Architecture)**.** *Suppose Assumptions 1 and 3 hold. Let* $\widehat{f}_{\text{DNN}}$ *be the deep ReLU network estimator defined by* (2.3)*, for either least squares* (2.1) *or logistic* (2.2) *loss. Then with probability at least* $1 - e^{-\gamma}$*,*

(a) $\|\widehat{f}_{\text{DNN}} - f_*\|^2_{L_2(X)} \leq C \left(\frac{W_n L_n \log U_n}{n} \log n + \frac{\log\log n + \gamma}{n} + \epsilon^2_{\text{DNN}} \right)$ *and*

(b) $\mathbb{E}_n \left[(\widehat{f}_{\text{DNN}} - f_*)^2 \right] \leq C \left(\frac{W_n L_n \log U_n}{n} \log n + \frac{\log\log n + \gamma}{n} + \epsilon^2_{\text{DNN}} \right)$,

for a universal constant $C > 0$ *independent of* n*.*

This result covers the general deep ReLU network problem defined in (2.3) for general feedforward architectures. The same comments as were made following Theorem 1 apply here as well: the same localization argument is used with the same benefits. We explicitly use this in the next two corollaries, where we exploit the allowed flexibility in controlling ϵ_{DNN} by stating results for particular architectures. The bound here is not directly applicable without specifying the network structure, which will determine both the variance portion (through W_n, L_n, and U_n) and the approximation ϵ_{DNN}. With these set, the bound becomes operational upon choosing γ, which can be optimized as desired, and this will immediately then yield a convergence rate.

Turning to special cases, we first show that the optimal rate of Stone (1982) can be attained, up to log factors. However, this relies on a rather artificial network structure, designated to approximate functions in a Sobolev space well, but without concern for practical implementation. Thus, while the following rate improves upon Theorem 1, we view this result as mainly of theoretical interest: establishing that (certain) deep ReLU networks are able to attain the optimal rate.

Corollary 1 (Optimal Rate)**.** *Suppose Assumptions 1 and 2 hold. Let* $\widehat{f}_{\text{OPT}}$ *solve* (2.3) *using the (deep and wide) network of Yarotsky (2017, Theorem 1), with* $W_n \asymp U_n \asymp n^{\frac{d}{2\beta+d}} \log n$ *and* $L_n \asymp \log n$*, the following hold with probability at least* $1 - e^{-\gamma}$*,*

(a) $\|\widehat{f}_{\mathrm{OPT}} - f_*\|^2_{L_2(X)} \leq C \cdot \left\{ n^{-\frac{2\beta}{2\beta+d}} \log^4 n + \frac{\log\log n + \gamma}{n} \right\}$ *and*

(b) $\mathbb{E}_n \left[(\widehat{f}_{\mathrm{OPT}} - f_*)^2 \right] \leq C \cdot \left\{ n^{-\frac{2\beta}{2\beta+d}} \log^4 n + \frac{\log\log n + \gamma}{n} \right\}$,

for a universal constant $C > 0$ independent of n.

Next, we turn to *very* deep networks that are very narrow, which have attracted substantial recent interest. Theorem 1 and Corollary 1 dealt with networks where the depth grows slowly with sample size and the width is polynomial in sample size. This matches the most common empirical practice, and is what we use in Section 4. However, it is possible to allow for networks of *fixed* width, provided the depth is sufficiently large. Using recent results (see also Mhaskar and Poggio, 2016b; Hanin, 2017) we can establish the following result for very deep MLPs.

Corollary 2 (Fixed Width Networks)**.** *Let the conditions of Theorem 1 hold, with $\beta = 1$ in Assumption 2. Then for the MLP of Yarotsky (2018, Theorem 1) (very deep and fixed width), with $H = 2d + 10$ and $L \asymp n^{\frac{d}{2(2+d)}}$, the following hold with probability at least $1 - e^{-\gamma}$,*

(a) $\|\widehat{f}_{\mathrm{MLP}} - f_*\|^2_{L_2(X)} \leq C \cdot \left\{ n^{-\frac{2}{2+d}} \log^2 n + \frac{\log\log n + \gamma}{n} \right\}$ *and*

(b) $\mathbb{E}_n \left[(\widehat{f}_{\mathrm{MLP}} - f_*)^2 \right] \leq C \cdot \left\{ n^{-\frac{2}{2+d}} \log^2 n + \frac{\log\log n + \gamma}{n} \right\}$,

for a universal constant $C > 0$ independent of n.

This result is again mainly of theoretical interest. The restriction to $\beta = 1$ here limits the potential applications of the result because, in practice, d will be large enough to render this rate, unlike those above, too slow for use in later inference procedures. In particular, if $d \geq 3$, the rate will not be sufficient for causal inference (in particular, the conditions of Theorem 3 fail to hold).

Remark 2. Finally, we note that although there has been a great deal of work in easing implementation (optimization and tuning) of deep nets, we have ignored this important aspect so far. Implementation can be a challenge in many settings, particularly when using non-standard architectures. See also Remark 1. Given the renewed interest in deep networks, this is an area of study already (Hartford et al., 2017; Polson and Rockova, 2018) and we expect this to continue and that implementations will rapidly evolve. This is perhaps another reason that Theorem 1 is, at the present time, the most practically useful. We will revisit this issue in our empirical illustration in Section 4. ⌟

3 Semiparametric Causal Inference

We now use the results above, in particular Theorem 1, coupled with results in the semiparametric literature, to deliver valid asymptotic inference for causal effects. For concreteness, we focus on two parameters of interest in our empirical application: average treatment effects and expected profits under different targeting policies. The former is very widely studied, and has served as a benchmark parameter in the study of inference following machine learning (see Section 1.1). The latter parameter shares many similar features and in many contexts is equally useful in decision-making. The novelty of our results is not in this semiparametric stage per se, but rather in delivering valid inference relying on deep neural networks for the first step estimation.

3.1 Average Treatment Effects

The estimation of average treatment effects is a well-studied problem, and we will give only a brief overview here. Recent reviews and further references are given by Belloni et al. (2017); Athey et al. (2017); Abadie and Cattaneo (2018). We consider the standard setup for program evaluation with observational data: we observe a sample of n units, each exposed to a binary treatment, and for each unit we observe a vector of pre-treatment covariates, $\boldsymbol{X} \in \mathbb{R}^d$, treatment status $T \in \{0, 1\}$, and a scalar post-treatment outcome Y. The observed outcome obeys $Y = TY(1) + (1-T)Y(0)$, where $Y(t)$ is the (potential) outcome under treatment status $t \in \{0, 1\}$. The "fundamental problem" is that only $Y(0)$ or $Y(1)$ is observed for each unit, never both. The parameter of interest is the average treatment effect

$$\tau = \mathbb{E}[Y(1) - Y(0)]. \tag{3.1}$$

In the context of our empirical example, the treatment is being mailed a catalog and the outcome is either a binary purchase decision or dollars spent. The average treatment effect, also referred to as "lift" in digital contexts, corresponds to the expected gain in revenue from an average individual receiving the catalog compared to the same person not receiving the catalog.

The crucial assumption under which τ is identified is selection on observables, also known as ignorability, unconfoundedness, missingness at random, or conditional independence, stated as follows. Let $p(\boldsymbol{x}) = \mathbb{P}[T = 1|\boldsymbol{X} = \boldsymbol{x}]$ denote the propensity score and $\mu_t(\boldsymbol{x}) = E[Y(t)|\boldsymbol{X} = \boldsymbol{x}]$, $t \in \{0, 1\}$ denote the two regression functions.

Assumption 4. *For $t \in \{0, 1\}$ and almost surely $\boldsymbol{X}$, $\mathbb{E}[Y(t)|T, \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y(t)|\boldsymbol{X} = \boldsymbol{x}]$ and $\bar{p} \leq p(\boldsymbol{x}) \leq 1 - \bar{p}$ for some $\bar{p} > 0$.*

This is the central assumption we maintain throughout. Beyond this, we will mostly need only regularity conditions.

Our results will apply immediately to other average treatment effect estimands, such as the treatment effect on the treated or multi-valued treatments; we focus on τ for concreteness. Different treatment effect estimands are reviewed by Lechner (2001), Cattaneo (2010), and others. Further, under selection on observables, treatment effects, missing data, measurement error, and data combination models are equivalent, and thus all our results apply immediately to those contexts. For reviews of these issues and Assumption 4 more broadly, see Chen et al. (2004); Tsiatis (2006); Heckman and Vytlacil (2007); Imbens and Wooldridge (2009).

3.2 Expected Profits

A closely related parameter of interest is the expected realized outcome. We will refer to this as the expected profit, for simplicity and in anticipation of our empirical application, but in general this may be interpreted as the total welfare from a treatment policy, or total health outcome in a medical context. The question of interest here is whether a change in the treatment policy would be beneficial in terms of increasing outcomes, and this is judged using observational data. Intuitively, the average treatment effect is the expected gain from treatment for the "next" person exposed to treatment, relative to if they had not been exposed. That is, it is the expected change in the outcome. Expected profit, on the other hand, is concerned with the total outcome, not the difference in outcomes. In the context of our empirical application, here we are interested in the probability of making a purchase decision or in total sales, rather than the change in each.

Profits are a causal parameter, and we can maintain the same set up as above, with the addition of a hypothetical treatment targeting strategy that is the object of evaluation. This is simply a rule that assigns a given consumer profile, determined by the covariates X, to treatment status:

that is, a known function (i.e., not estimated from the sample) $s(x) : \operatorname{supp}\{\boldsymbol{X}\} \to \{0, 1\}$. Note well that this is *not* necessarily the observed treatment: $s(\boldsymbol{x}_i) \neq t_i$. The policy maker may wish to evaluate the gain from targeting only a certain subset of customers, a price discrimination strategy, or comparisons of different such policies.

The base parameter of interest is the profit from a fixed policy, given by

$$\pi(s) = \mathbb{E}\big[s(\boldsymbol{X})Y(1) + (1 - s(\boldsymbol{X}))\,Y(0)\big], \tag{3.2}$$

where we make explicit the dependence on the policy $s(\cdot)$. Compare to Equation (3.1) and recall that the *observed* outcome obeys $Y = TY(1) + (1-T)Y(0)$. Whereas τ is the gain in assigning the next person to treatment and is given by the difference in potential outcomes, $\pi(s)$ is the expected outcome that would be observed for the next person if the treatment rule were $s(\boldsymbol{x})$ as opposed to the T which generated the data.

A natural question is whether one targeting strategy, say $s_a(\boldsymbol{x})$, is superior to another, or to a status quo policy, say $s_b(\boldsymbol{x})$. This amounts to testing the hypothesis $H_0 : \pi(s_a) \geq \pi(s_b)$. To evaluate this, we can study the difference in expected profits, which amounts to

$$\pi(s_a, s_b) = \pi(s_a) - \pi(s_b) = \mathbb{E}\big[(s_a(\boldsymbol{X}) - s_b(\boldsymbol{X}))Y(1) + (s_b(\boldsymbol{X}) - s_a(\boldsymbol{X}))\,Y(0)\big]. \tag{3.3}$$

Assumption 4 provides identification for $\pi(s)$ and $\pi(s_a, s_b)$, arguing analogously as for τ. Moreover, notice that $\pi(s_a, s_b) = \mathbb{E}[(s_a(\boldsymbol{X}) - s_b(\boldsymbol{X}))(Y(1) - Y(0))] = \mathbb{E}[(s_a(\boldsymbol{X}) - s_b(\boldsymbol{X}))\tau(\boldsymbol{X})]$, where $\tau(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]$ is the conditional average treatment effect. The latter form makes clear that only those differently treated, of course, impact the evaluation of s_a compared to s_b. The strategy s_a will be superior if, on average, it targets those with a higher individual treatment effect, $\tau(\boldsymbol{x})$. Other parameters of interest can be considered. An obvious example is evaluating an estimated optimal treatment policy that has been estimated from auxiliary data or via sample splitting.

3.3 Asymptotic Results

Our estimation of τ, $\pi(s)$, and $\pi(s_a, s_b)$ will utilize doubly robust estimators. These estimators, based in this context on sample analogues of the efficient influence function (Hahn, 1998), are known as doubly robust as they remain consistent if either the regression functions or the propensity score are correctly specified (Robins et al., 1994, 1995). Our use here follows the recent literature in econometrics showing that the double robustness implies valid inference under weaker conditions on the first step nonparametric estimates (Belloni et al., 2014; Farrell, 2015; Chernozhukov et al., 2018a). Indeed, it is these conditions that we verify using the results of Theorem 1.

To define these estimators, suppose we have a sample $\{\boldsymbol{z}_i = (y_i, t_i, \boldsymbol{x}_i')'\}_{i=1}^n$ from $\boldsymbol{Z} = (Y, T, \boldsymbol{X}')'$. Then for $t \in \{0, 1\}$ define

$$\psi_t(\boldsymbol{z}_i) = \frac{\mathbb{1}\{t_i = t\}(y_i - \mu_t(\boldsymbol{x}_i))}{\mathbb{P}[T = t \mid X = \boldsymbol{x}_i]} + \mu_t(\boldsymbol{x}_i) \tag{3.4}$$

and its sample analogue

$$\hat{\psi}_t(\boldsymbol{z}_i) = \frac{\mathbb{1}\{t_i = t\}(y_i - \hat{\mu}_t(\boldsymbol{x}_i))}{\hat{\mathbb{P}}[T = t \mid X = \boldsymbol{x}_i]} + \hat{\mu}_t(\boldsymbol{x}_i), \tag{3.5}$$

where $\hat{\mathbb{P}}[T = t \mid X = \boldsymbol{x}_i] = \hat{p}(\boldsymbol{x}_i)$ for $t = 1$ and $1 - \hat{p}(\boldsymbol{x}_i)$ for $t = 0$. Then we define the following,

$$\begin{aligned} \hat{\tau} &= \mathbb{E}_n \left[\hat{\psi}_1(\boldsymbol{z}_i) - \hat{\psi}_0(\boldsymbol{z}_i) \right], \\ \hat{\pi}(s) &= \mathbb{E}_n \left[s(\boldsymbol{x}_i)\hat{\psi}_1(\boldsymbol{z}_i) + (1 - s(\boldsymbol{x}_i))\hat{\psi}_0(\boldsymbol{z}_i) \right], \\ \hat{\pi}(s_a, s_b) &= \mathbb{E}_n \left[[s_a(\boldsymbol{x}_i) - s_b(\boldsymbol{x}_i)]\hat{\psi}_1(\boldsymbol{z}_i) - [s_a(\boldsymbol{x}_i) - s_b(\boldsymbol{x}_i)]\hat{\psi}_0(\boldsymbol{z}_i) \right]. \end{aligned} \tag{3.6}$$

The estimator $\hat{\tau}$ is exactly the doubly/locally robust estimator of the average treatment effect that is standard in the literature. The estimators for profits can be thought of as the doubly robust version of the constructs described in Hitsch and Misra (2018). Furthermore, to add a per-unit cost of treatment/targeting c and a margin m, simply replace ψ_1 with $m\psi_1 - c$ and ψ_0 with $m\psi_0$.

For the first stage estimates appearing in (3.5) we use our results on deep nets, and Theorem 1 in particular. Specifically, the estimated propensity score, $\hat{p}(\boldsymbol{x})$, is the estimate that results from solving (2.3), with the MLP architecture, for the logistic loss (2.2) with T as the outcome. Similarly, for each status $t \in \{0, 1\}$, let $\hat{\mu}_t(\boldsymbol{x})$ be the deep-MLP estimate of $f_*(\boldsymbol{x}) = \mathbb{E}[Y | T = t, \boldsymbol{X} = \boldsymbol{x}]$, solving

(2.3) for least squares loss, (2.1), with outcome Y, using only observations with $t_i = t$. In the case of binary outcomes, the latter are replaced by logistic regressions.

We then obtain inference using the following results, essentially taken from Farrell (2015). Similar results are given by Belloni et al. (2014, 2017, 2018). All of these provide high-level conditions for valid inference, and none verify these for deep nets as we do here. We opt to state a generic results for $\psi_t(\boldsymbol{z}_i)$, given their shared structure, which immediately covers τ, $\pi(s)$, and $\pi(s_a, s_b)$, rather than repeat the same conditions for each estimand.

Theorem 3. *Suppose that $\{\boldsymbol{z}_i = (y_i, t_i, \boldsymbol{x}_i')'\}_{i=1}^n$ are i.i.d. obeying Assumption 4 and the conditions Theorem 1 hold with $\beta > d$. Further assume that, for $t \in \{0,1\}$, $\mathbb{E}[(s(\boldsymbol{X})\psi_t(\boldsymbol{Z}))^2 | \boldsymbol{X}]$ is bounded away from zero and $\mathbb{E}[(s(\boldsymbol{X})\psi_t(\boldsymbol{Z}))^{4+\delta} | \boldsymbol{X}]$ is bounded from some $\delta > 0$. Then the deep ReLU network estimators defined above obey the following, for $t \in \{0,1\}$,*

- **(a)** $\mathbb{E}_n[(\hat{p}(\boldsymbol{x}_i) - p(\boldsymbol{x}_i))^2] = o_P(1)$ *and* $\mathbb{E}_n\left[(\hat{\mu}_t(\boldsymbol{x}_i) - \mu_t(\boldsymbol{x}_i))^2\right] = o_P(1)$,
- **(b)** $\mathbb{E}_n[(\hat{\mu}_t(\boldsymbol{x}_i) - \mu_t(\boldsymbol{x}_i))^2]^{1/2}\mathbb{E}_n[(\hat{p}(\boldsymbol{x}_i) - p(\boldsymbol{x}_i))^2]^{1/2} = o_P(n^{-1/2})$, *and*
- **(c)** $\mathbb{E}_n[(\hat{\mu}_t(\boldsymbol{x}_i) - \mu_t(\boldsymbol{x}_i))(1 - \mathbb{1}\{t_i = t\}/\mathbb{P}[T = t | \boldsymbol{X} = \boldsymbol{x}_i])] = o_P(n^{-1/2})$,

and therefore for a given $s(\boldsymbol{x})$ and $t \in \{0,1\}$,

$$\sqrt{n}\mathbb{E}_n\left[s(\boldsymbol{x}_i)\hat{\psi}_t(\boldsymbol{z}_i) - s(\boldsymbol{x}_i)\psi_t(\boldsymbol{z}_i)\right] = o_P(1) \qquad \textit{and} \qquad \frac{\mathbb{E}_n[(s(\boldsymbol{x}_i)\hat{\psi}_t(\boldsymbol{z}_i))^2]}{\mathbb{E}_n[(s(\boldsymbol{x}_i)\psi_t(\boldsymbol{z}_i))^2]} = o_P(1).$$

It is immediate from this result that the estimators $\hat{\tau}$, $\hat{\pi}(s)$, and $\hat{\pi}(s_a, s_b)$ are asymptotically Normal. The asymptotic variance can be estimated by simply replacing the sample first moments of (3.6) with second moments. That is, looking at $\hat{\pi}(s)$ for example,

$$\sqrt{n}\hat{\Sigma}^{-1/2}\left(\hat{\pi}(s) - \pi(s)\right) \xrightarrow{d} \mathcal{N}(0,1), \quad \text{with} \quad \hat{\Sigma} = \mathbb{E}_n\left[\left(s(\boldsymbol{x}_i)\hat{\psi}_1(\boldsymbol{z}_i) + (1 - s(\boldsymbol{x}_i))\hat{\psi}_0(\boldsymbol{z}_i)\right)^2\right] - \hat{\pi}(s)^2.$$

The others are similar. This result shows exactly how the deep ReLU networks deliver valid asymptotic inference for our parameters of interest. Theorem 1 proves that the nonparametric estimates converge sufficiently fast, as formalized by conditions (a), (b) and (c), enabling feasible semiparametric efficient inference. In general, these are implied by, but may be weaker than, the requirement of that the first step estimates converge faster than $n^{-1/4}$, which our results yield for deep ReLU nets. The first is a mild consistency requirement. The second requires a rate, but on the product of the two estimates, which can be satisfied under weaker conditions. Finally, the third condition is the strongest. Intuitively, this condition arises from a "leave-in" type remainder, and as such, it can be weakened using sample splitting (Newey and Robins, 2017). We opt to maintain (c) exactly because deep nets are not amenable to either simple leave-one-out forms (as are, e.g., classical kernel regression) or to sample splitting; being a data hungry method the gain in rate requirements may not be worth the price paid in constants. We use our localization approach to verify (c), see Lemma 11, we may also be of interest in future applications of second-step inference using machine learning methods.

Theorem 3 can be generalized straightforwardly to both cover other estimands and to hold uniformly over data-generating processes. Any estimand with a locally/doubly robust estimator that depends only on conditional expectations can in principle be covered by our results. In particular, our results can be used as an ingredient in verifying the conditions of Chernozhukov et al. (2018b, Section 7), who treat more general semiparametric estimands using local robustness, sometimes relying on sample splitting as discussed above. When locally robust estimators are not available we can obtain similar results (that is, under weak rate restrictions) by relying on sample splitting. This can be useful for more complex/structured cases, where deriving locally robust moment conditions is prohibitive. Moreover, under appropriate assumptions, such results and Theorem 3 more specifically, will hold uniformly over distributions, as discussed by, e.g. Belloni et al. (2017).

3.4 Inference Under Randomization

Our analysis thus far has focused on observational data, but it is worth spelling out results for randomized experiments. This is particularly important in the Internet age, where experimentation is common, vast amounts of data are available, and effects are often small in magnitude (Taddy et al., 2015). Indeed, our empirical illustration, detailed in the next section, stems from an experiment with 300,000 units and hundreds of covariates. When treatment is randomized, inference on τ and π can be done easily and directly using the difference in means between the treatment and control groups for the average treatment effect or the corresponding weighted sum for profits. Thus it may seem that machine learning methods are mismatched to this type of data. However, when pre-treatment covariates are available they can be used to increase efficiency (Hahn, 2004).

We will focus on the simple situation of a purely randomized binary treatment, but our results can be extended naturally to other randomization schemes. We formalize this with the following.

Assumption 5 (Randomized Treatment)**.** *T is independent of $Y(0)$, $Y(1)$, and X, and is distributed Bernoulli with parameter p^*, such that $\bar{p} \leq p^* \leq 1 - \bar{p}$ for some $\bar{p} > 0$.*

Under this assumption, the obvious simplification is that the propensity score need not be estimated using the covariates, but can be replaced with the (still nonparametric) sample frequency: $\hat{p}(\boldsymbol{x}_i) \equiv \hat{p} = \mathbb{E}_n[t_i]$. This is plugged into Equation (3.6) and estimation and inference proceeds as above. For valid inference, only rate conditions on the regression functions $\hat{\mu}_t(x)$ are needed. Further, conditions (a) and (b) of Theorem 3 collapse, as $\hat{p}$ is root-n consistent, leaving only condition (c) to be verified. We collect this into the following result, which is a trivial corollary of Theorem 3.

Corollary 3. *Let the conditions of Theorem 3 hold with Assumption 5 in place of Assumption 4. Then deep ReLU network estimators obey*

(a′) $\mathbb{E}_n\left[(\hat{\mu}_t(\boldsymbol{x}_i) - \mu_t(\boldsymbol{x}_i))^2\right] = o_P(1)$ *and*

(c′) $\mathbb{E}_n[(\hat{\mu}_t(\boldsymbol{x}_i) - \mu_t(\boldsymbol{x}_i))(1 - \mathbb{1}\{t_i = t\}/\mathbb{P}[T = t|\boldsymbol{X} = \boldsymbol{x}_i])] = o_P(n^{-1/2})$

and the conclusions of Theorem 3 hold.

This is the result we make use of in our empirical illustration.

4 Empirical Application

To illustrate our results, Theorems 1 and 3 in particular, we study, from a marketing point of view, a randomized experiment from a large US retailer of various consumer products. The outcome of interest is consumer spending and the treatment is a catalog mailing. The firm sells directly to the customer (as opposed to via retailers) using a variety of channels such as the web and mail. The data consists of nearly three hundred thousand (292,657) consumers chosen at random from the retailer's database. Of these, 2/3 were randomly chosen to receive a catalog, and in addition to treatment status, we observe roughly one hundred fifty covariates, including demographics, past purchase behaviors, interactions with the firm, and other relevant information. For more detail on this data, as well as a complete discussion of the decision making issues, we refer the reader to Hitsch and Misra (2018) (we use the 2015 sample). That paper studied various estimators, both traditional and modern, of average and heterogeneous treatment effects. Importantly, they did not consider neural networks. Our results suggest that deep nets are at least as good as (and sometimes better) that the best methods found by Hitsch and Misra (2018).

In general, a key element of a firm's toolkit is the design and implementation of targeted marketing instruments. These instruments, aiming to induce demand, often contain advertising and informational content about the firms offerings. The targeting aspect thus boils down to the selection of which particular customers should be sent the material. This is a particularly important decision since the costs of creation and dissemination the material can accumulate rapidly, particularly over a large customer base. For a typical retailer engaging in direct marketing the costs of sending out a catalog (a print booklet with the firms products for sale) can be cost close to a dollar per targeted customer. With millions of catalogs being sent out, the cost of a typical campaign is quite high.

Given these expenses an important problem for firms is ascertaining the causal effects of such targeted mailing, and then using these effects to evaluate potential targeting strategies. At a high level, this approach is very similar to personalized modern medicine where treatments have to be targeted. In these contexts, both the treatment and the targeting can be extremely costly, and thus careful assessment of $\pi(s)$ (interpreted here as welfare) is crucial for decision making.

The outcome of interest for the firm is customer spend. This is the total amount of money that a given customer spends on purchases of the firm's products, within a specified time window. For the experiment in question the firm used a window of three months, and aggregated sales from all available purchase channels including phone, mail, and the web. In our data 6.2% of customers made a purchase. Overall mean spending is \$7.31; average spending conditional on buying is \$117.7, with a standard deviation of \$132.44. Figure 3 displays the complete density of spending conditional on a purchase, which is quite skewed. The idea then is to examine the incremental effect that the catalog had on this spending metric. Table 1 presents summary statistics for the outcome and treatment.

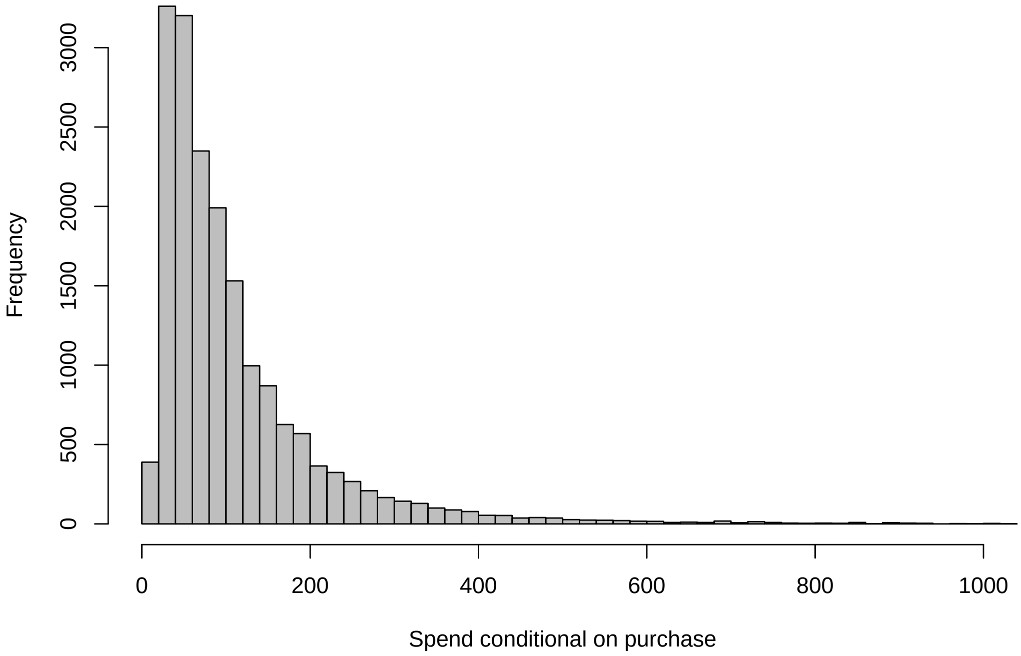


Figure 3: Spend Conditional on Purchase

4.1 Implementation Details

We estimated deep neural nets under a variety of architecture choices. For our particular empirical application simpler architectures performed better than more complex ones. In what follows we present eight examples and focus on one particular architecture to compute various statistics and tests to illustrate the use of the theory developed above. All computation was done using TensorFlow™.

For treatment effect and profit estimation we follow Equations (3.5) and (3.6). Because treatment

Table 1: Summary Statistics

	Mean	SD	N
Purchase	0.062	0.24	292657
Spend	7.311	43.55	292657
Spend Conditional on Purchase	117.730	132.44	18174
Treatment	0.669	0.47	292657

is randomized, we apply Corollary 3, and thus, only require estimates of the regression functions $\mu_t(\boldsymbol{x}) = E[Y(t)|\boldsymbol{X} = \boldsymbol{x}]$, $t \in \{0,1\}$. An important implementation detail, from a computation point of view (recall Remark 2) is that we will estimate μ_0 and μ_1 jointly (results from separate estimation are available). To be precise, recalling Equations (2.1) and (2.3), we solve

$$\begin{pmatrix} \hat{\mu}_0(\boldsymbol{x}) \\ \hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x}) \end{pmatrix} := \underset{\tilde{\mu}_0, \tilde{\tau}}{\arg\min} \sum_{i=1}^{n} \frac{1}{2}\Big(y_i - \tilde{\mu}_0(\boldsymbol{x}_i) - \tilde{\tau}(\boldsymbol{x}_i)t_i\Big)^2$$

where the minimization is over the relevant network architecture. In the context of our empirical example y_i is the costumer's spending, $\boldsymbol{x}_i$ are her characteristics, and t_i indicates receipt of a catalog. In this format, $\mu_0(\boldsymbol{x}_i)$ reflects base spending and $\tau(\boldsymbol{x}) = \mu_1(\boldsymbol{x}_i) - \mu_0(\boldsymbol{x}_i)$ is the conditional average treatment effect of the catalog mailing. In our application, this joint estimation outperforms separately estimating each $\mu_t(\boldsymbol{x})$ on the respective samples (though these two approaches are equivalent theoretically).

The details of the eight deep net architectures are presented in Table 2. See Section 2.1 for an introduction to the terminology and network construction. Most yielded similar results, both in terms of fit and final estimates. A key measure of fit reported in the final column of the table is the portion of $\hat{\tau}(\boldsymbol{x}_i)$ that were negative. As argued by Hitsch and Misra (2018), it is implausible under standard marketing or economic theory that receipt of a catalog causes lower purchasing. On this metric of fit, deep nets perform as well as, and sometimes better than, the best methods found by Hitsch and Misra (2018): Causal KNN with Treatment Effect Projections (detailed therein) or Causal Forests (Wager and Athey, 2018). Figure 4 shows the distribution of $\hat{\tau}(\boldsymbol{x}_i)$ across customers for each of the eight architectures. While there are differences in the shapes of the densities, the mean and variance estimates are nonetheless quite similar.

4.2 Results

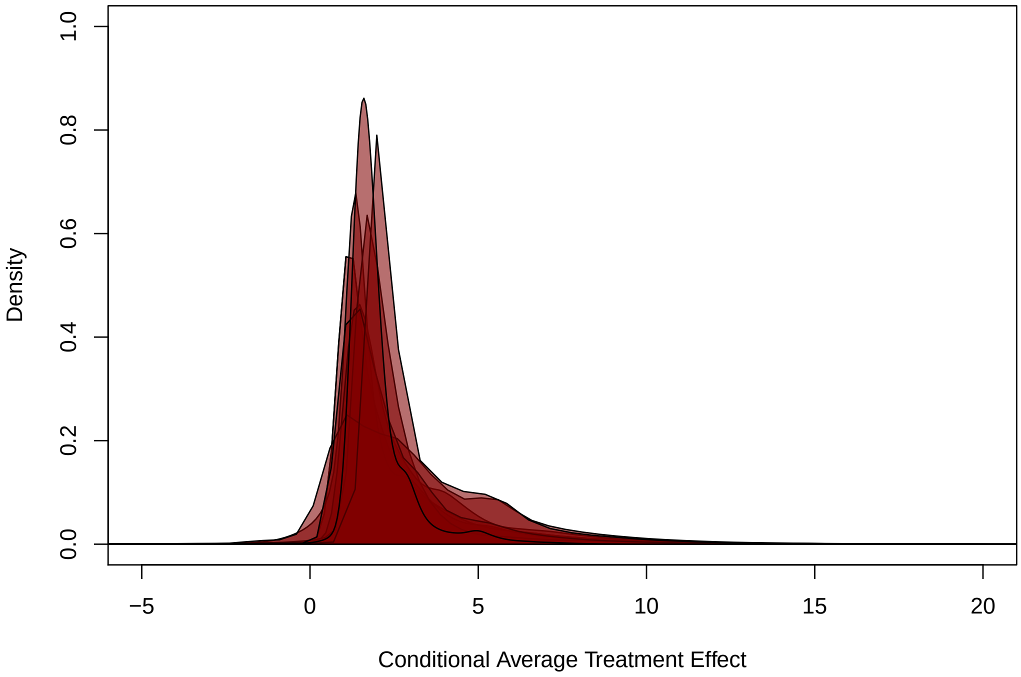
We present results for treatment effects, profits, and targeting policy evaluations. Table 3 shows the estimates of the average treatment effect from the eight network architectures along with their respective 95% confidence intervals. These results are constructed following Section 3, using

Table 2: Deep Network Architectures

Architecture	Learning Rate	Widths $[H_1, H_2, ...]$	Total Parameters (W)	Validation Loss	Training Loss	$\mathbb{P}_n[\hat{\tau}(\boldsymbol{x}_i) < 0]$
1	0.0003	[60]	8702	1405.62	1748.91	0.0014
2	0.0003	[100]	14502	1406.48	1751.87	0.0251
3	0.0001	[30, 20]	4952	1408.22	1751.20	0.0072
4	0.0009	[30, 10]	4622	1408.56	1751.62	0.0138
5	0.0003	[30, 30]	5282	1403.57	1738.59	0.0226
6	0.0003	[30, 30]	5282	1408.57	1755.28	0.0066
7	0.0003	[100, 30, 20]	17992	1408.62	1751.52	0.0103
8	0.00005	[80, 30, 20]	14532	1413.70	1756.93	0.0002

Notes: All networks use the ReLU activation function. The width of each layer is shown, e.g. Architecture 3 consists of two layers, with 30 and 20 hidden units respectively. The final column shows the portion of estimated individual treatment effects below zero.

Figure 4: Conditional Average Treatment Effects Across Architectures



Equations (3.5) and (3.6) in particular, and valid by Corollary 3. All eight specifications yield quite similar results. Furthermore, because this is an experiment, we can compare to the standard unadjusted difference in means, which yields an average treatment effect of 2.632 and associated 95% interval $[2.303, 2.960]$.

Table 3: Average Treatment Effect Estimates and 95% Confidence Intervals

Architecture	Average Treatment Effect ($\hat{\tau}$)	95% Confidence Interval
1	2.606	[2.273 , 2.932]
2	2.577	[2.252 , 2.901]
3	2.547	[2.223 , 2.872]
4	2.488	[2.160 , 2.817]
5	2.459	[2.127 , 2.791]
6	2.430	[2.093 , 2.767]
7	2.400	[2.057 , 2.744]
8	2.371	[2.021 , 2.721]

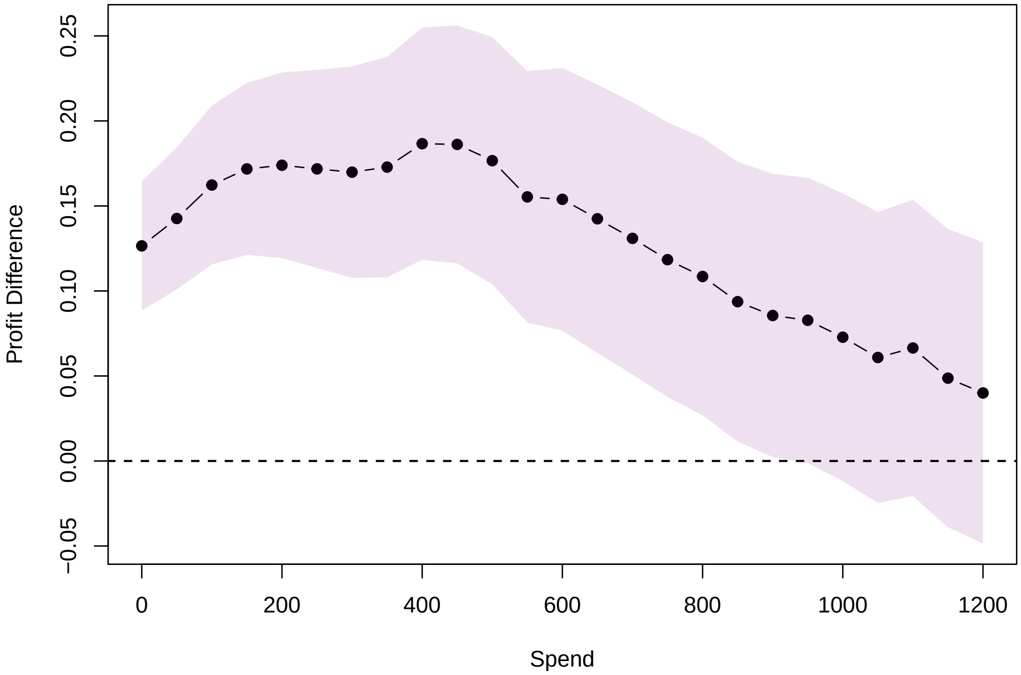
Table 4: Counterfactual Profits from Three Targeting Strategies

	Never Treat		**Blanket Treatment**		**Loyalty Policy**	
Architecture	$\hat{\pi}(s)$	95% CI	$\hat{\pi}(s)$	95% CI	$\hat{\pi}(s)$	95% CI
1	2.016	[1.923 , 2.110]	2.234	[2.162 , 2.306]	2.367	[2.292 , 2.443]
2	2.022	[1.929 , 2.114]	2.229	[2.157 , 2.301]	2.363	[2.288 , 2.438]
3	2.027	[1.934 , 2.120]	2.224	[2.152 , 2.296]	2.358	[2.283 , 2.434]
4	2.037	[1.944 , 2.130]	2.213	[2.140 , 2.286]	2.350	[2.274 , 2.425]
5	2.043	[1.950 , 2.136]	2.208	[2.135 , 2.281]	2.345	[2.269 , 2.422]
6	2.048	[1.954 , 2.142]	2.202	[2.128 , 2.277]	2.341	[2.263 , 2.418]
7	2.053	[1.959 , 2.148]	2.197	[2.122 , 2.272]	2.336	[2.258 , 2.414]
8	2.059	[1.963 , 2.154]	2.192	[2.116 , 2.268]	2.332	[2.253 , 2.411]

Turning to expected profits, we estimate $\pi(s) = \mathbb{E}\big[s(\boldsymbol{X})(mY(1) - c) + (1 - s(\boldsymbol{X}))\,mY(0)\big]$, adding a profit margin m and a mailing cost c to (3.2) (our NDA with the firm forbids revealing m and c). We consider three different counterfactual policies $s(\boldsymbol{x})$: (i) *never* treat, $s(\boldsymbol{x}) \equiv 0$; (ii) a *blanket* treatment, $s(\boldsymbol{x}) \equiv 1$; (iii) a *loyalty* policy, $s(\boldsymbol{x}_i) = 1$ only for those who had purchased in the prior calendar year. Results are shown in Table 4. There is close agreement among the eight architectures both numerically and substantially: it is clear that profits from the three policies are ordered as $\pi(\text{never}) < \pi(\text{blanket}) < \pi(\text{loyalty})$.

To explore further, we focus on specification #3 and study further subpopulation treatment targeting strategies. (The other architectures yield similar results, so we omit them.) Architecture #3 has depth $L = 2$ with widths $H_1 = 30$ and $H_2 = 20$. The learning rate was set at 0.0001 and the specification had a total of 4,952 parameters. For this architecture, recalling Remark 1, we added dropout for the second layer with a fixed probability of 1/2. Using this architecture, we compare the blanket strategy to targeting customers with spend of at least $\bar{y}$ dollars in the prior

Figure 5: Expected Profits from Threshold Targeting Based on Prior Year Spend



calendar year. Figure 5 presents the results for $0 \leq \bar{y} \leq 1200$. The black dots show the difference $\{\pi(\text{spend} > \bar{y}) - \pi(\text{blanket})\}$ and the shaded region gives *pointwise* 95% confidence bands. We see that there is a significant different between various choices of $\bar{y}$. Initially, targeting customers with higher spend yields higher profits, as would be expected, but this effect diminishes beyond a certain $\bar{y}$, roughly \$500, as fewer and fewer are targeted.

5 Conclusion

We have demonstrated new rates of convergence for deep neural networks. Our results handle different network architectures, but importantly cover the modern standard practice of fully-connected, feedfoward networks, that is, multi-layer perceptrons, and use of the ReLU activation function. The convergence rates are sufficiently fast to deliver semiparametric inference, where we have focused on treatment effects and profits. To the best of our knowledge, this is the first inference result using deep nets. Our results cover nonparametric conditional expectations, both least squares and logistic loss, and are thus widely applicable. For some estimands, it may be crucial to estimate the density as well, and this problem can be challenging in high dimensions. Deep nets, in the form of Generative Adversarial Networks are a promising tool for density estimation; see Liang (2017) for recent results. Research into these further applications and structures is underway.

6 References

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

Anthony, M., and Bartlett, P. L. (1999), *Neural Network Learning: Theoretical Foundations*, Campbridge University Press.

Athey, S., Imbens, G., Pham, T., and Wager, S. (2017), "Estimating average treatment effects: Supplementary analyses and remaining challenges," *American Economic Review: Papers & Proceeding*, 107, 278–81.

Athey, S., Imbens, G. W., and Wager, S. (2016), "Approximate residual balancing: De-biased inference of average treatment effects in high dimensions," *arxiv:1604.07125, Journal of the Royal Statistical Society, Series B,* forthcoming.

Barron, A. R. (1993), "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information theory*, 39, 930–945.

Bartlett, P. L., Bousquet, O., Mendelson, S. et al. (2005), "Local rademacher complexities," *The Annals of Statistics*, 33, 1497–1537.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2017), "Nearly-tight VC-dimension bounds for piecewise linear neural networks," in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2017)*.

Bauer, B., and Kohler, M. (2017), "On Deep Learning as a remedy for the curse of dimensionality in nonparametric regression," Technical report, Technical report.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018), "High-Dimensional Econometrics and Generalized GMM," *arXiv preprint arXiv:1806.01888*.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

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

Belloni, A., Chernozhukov, V., and Wang, L. (2011), "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, 98, 791–806.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of LASSO and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732.

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

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, eds. J. Heckman and E. Leamer, Vol. 6B of *Handbook of Econometrics*, chapter 76, Elsevier.

Chen, X., Hong, H., and Tarozzi, A. (2004), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," *Cowles Foundation Discussion Paper No. 1644*.

Chen, X., and Shen, X. (1998), "Sieve extremum estimates for weakly dependent data," *Econometrica*, 66, 289–314.

Chen, X., and White, H. (1999), "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2018b), "Locally Robust Semiparametric Estimation," *arXiv:1608.00033*.

Daniely, A. (2017), "Depth separation for neural networks," *arXiv preprint arXiv:1702.08489*.

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.

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

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

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

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

Hansen, C., Kozbur, D., and Misra, S. (2017), "Targeted Undersmoothing," *arXiv:1706.07328*.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017), "Deep iv: A flexible approach for counterfactual prediction," in *International Conference on Machine Learning*, pp. 1414–1423.

He, K., Zhang, X., Ren, S., and Sun, J. (2016), "Identity mappings in deep residual networks," in *European conference on computer vision*, Springer, pp. 630–645.

Heckman, J., and Vytlacil, E. J. (2007), "Econometric Evaluation of Social Programs, Part I," in *Handbook of Econometrics, vol. VIB*, eds. J. Heckman and E. Leamer, Elsevier Science B.V., pp. 4780–4874.

Hitsch, G. J., and Misra, S. (2018), "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation," *SSRN preprint 3111957*.

Hornik, K., Stinchcombe, M., and White, H. (1989), "Multilayer feedforward networks are universal approximators," *Neural networks*, 2, 359–366.

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Javanmard, A., and Montanari, A. (2014), "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, 15, 2869–2909.

Johansson, F., Shalit, U., and Sontag, D. (2016), "Learning representations for counterfactual inference," in *International Conference on Machine Learning*, pp. 3020–3029.

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

Koltchinskii, V. (2006), "Local Rademacher complexities and oracle inequalities in risk minimization," *The Annals of Statistics*, 34, 2593–2656.

Koltchinskii, V., and Panchenko, D. (2000), "Rademacher processes and bounding the risk of function learning," in *High dimensional probability II*, Springer, pp. 443–457.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012), "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105.

Lechner, M. (2001), "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," in *Econometric Evaluations of Active Labor Market Policies*, eds. M. Lechner and E. Pfeiffer, Heidelberg: Physica, pp. 43–58.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998), "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 86, 2278–2324.

Liang, T. (2017), "How Well Can Generative Adversarial Networks (GAN) Learn Densities: A Nonparametric View," *arXiv preprint arXiv:1712.08244*.

Liang, T., Rakhlin, A., and Sridharan, K. (2015), "Learning with square loss: Localization through offset Rademacher complexity," in *Conference on Learning Theory*, pp. 1260–1285.

Mendelson, S. (2003), "A few notes on statistical learning theory," in *Advanced lectures on machine learning*, Springer, pp. 1–40.

Mhaskar, H., and Poggio, T. (2016a), "Deep vs. shallow networks: An approximation theory perspective," *arXiv preprint arXiv:1608.03287*.

Mhaskar, H. N., and Poggio, T. (2016b), "Deep vs. shallow networks: An approximation theory perspective," *Analysis and Applications*, 14, 829–848.

Nair, V., and Hinton, G. E. (2010), "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Newey, W. K., and Robins, J. M. (2017), "Cross-fitting and fast remainder rates for semiparametric estimation," *arXiv preprint arXiv:1801.09138*.

Polson, N., and Rockova, V. (2018), "Posterior Concentration for Sparse Deep Learning," *arXiv preprint arXiv:1803.09138*.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. (2017), "On the Expressive Power of Deep Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, eds. D. Precup and Y. W. Teh, Vol. 70 of *Proceedings of Machine Learning Research*, International Convention Centre, Sydney, Australia: PMLR, pp. 2847–2854.

Robins, J., Li, L., Mukherjee, R., Tchetgen, E., and van der Vaart, A. (2017), "Minimax Estimation of a Functional on a Structured High-Dimensional Model," *The Annals of Statistics*, 45, 1951–1987.

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008), "Higher order influence functions and minimax estimation of nonlinear functionals," in *Probability and Statistics: Essays in Honor of David A. Freedman*, eds. D. Nolan and T. Speed, Vol. 2, Beachwood, Ohio, USA: Institute of Mathematical Statistics.

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

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

Safran, I., and Shamir, O. (2016), "Depth separation in relu networks for approximating smooth non-linear functions," *arXiv preprint arXiv:1610.09887*.

Schmidt-Hieber, J. (2017), "Nonparametric regression using deep neural networks with ReLU activation function," *arXiv preprint arXiv:1708.06633*.

Shalit, U., Johansson, F. D., and Sontag, D. (2017), "Estimating individual treatment effect: generalization bounds and algorithms," *arXiv preprint arXiv:1606.03976*.

Stone, C. J. (1982), "Optimal global rates of convergence for nonparametric regression," *The annals of statistics*, 1040–1053.

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2015), "A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation," *Arxiv preprint arXiv:1412.8563*.

Telgarsky, M. (2016), "Benefits of depth in neural networks," *arXiv preprint arXiv:1602.04485*.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer.

van de Geer, S., Buhlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202.

van der Laan, M., and Rose, S. (2001), *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer-Verlag.

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association,* forthcoming.

Westreich, D., Lessler, J., and Funk, M. J. (2010), "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression," *Journal of clinical epidemiology*, 63, 826–833.

White, H. (1989), "Learning in artificial neural networks: A statistical perspective," *Neural computation*, 1, 425–464.

——— (1992), *Artificial neural networks: approximation and learning theory*, Blackwell Publishers, Inc.

Yarotsky, D. (2017), "Error bounds for approximations with deep ReLU networks," *Neural Networks*, 94, 103–114.

Yarotsky, D. (2018), "Optimal approximation of continuous functions by very deep ReLU networks," *arXiv preprint arXiv:1802.03620.*

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016), "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530.*

A Proof of Theorems 1 and 2

In this section we provide a proof of Theorems 1 and 2, our main theoretical results for deep ReLU networks. The proof proceeds in several steps. We first give the main breakdown and bound the bias (approximation error) term. We then turn our attention to the empirical process term, to which we apply our localization. Much of the proof uses a generic architecture, and thus pertains to both results. We will specialize the architecture to the multi-layer perceptron only when needed later on. Other special cases and related results are covered in Section B. Supporting Lemmas are stated in Section C.

For notational simplicity we will denote $\widehat{f}_{\text{DNN}} := \hat{f}$, see (2.3), and $\epsilon_{\text{DNN}} := \epsilon_n$, see Assumption 3. As we are simultaneously consider Theorems 1 and 2, the generic notation DNN will be used throughout.

A.1 Main Decomposition and Bias Term

Referring to Assumption 3, define the best approximation realized by the deep ReLU network class $\mathcal{F}_{\text{DNN}}$ as

$$f_n := \mathop{\arg\min}_{\substack{f \in \mathcal{F}_{\text{DNN}} \\ \|f\|_\infty \leq 2M}} \|f - f_*\|_\infty .$$

By definition, $\epsilon_n := \epsilon_{\text{DNN}} := \|f_n - f_*\|_\infty$.

Recalling the optimality of the estimator in (2.3), we know, as both f_n and $\hat{f}$ are in $\mathcal{F}_{\text{DNN}}$, that

$$-\mathbb{E}_n[\ell(\hat{f}, \boldsymbol{z})] + \mathbb{E}_n[\ell(f_n, \boldsymbol{z})] \geq 0.$$

This result does not hold for f_* in place of f_n, because $f_* \notin \mathcal{F}_{\text{DNN}}$. Using the above display and Lemma 9 (which does not hold with f_n in place of f_* therein), we obtain

$$\begin{aligned} c_2 \|\hat{f} - f_*\|^2_{L_2(X)} &\leq \mathbb{E}[\ell(\hat{f}, \boldsymbol{z})] - \mathbb{E}[\ell(f_*, \boldsymbol{z})] \\ &\leq \mathbb{E}[\ell(\hat{f}, \boldsymbol{z})] - \mathbb{E}[\ell(f_*, \boldsymbol{z})] - \mathbb{E}_n[\ell(\hat{f}, \boldsymbol{z})] + \mathbb{E}_n[\ell(f_n, \boldsymbol{z})] \\ &= \mathbb{E}\left[\ell(\hat{f}, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right] - \mathbb{E}_n\left[\ell(\hat{f}, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right] + \mathbb{E}_n\left[\ell(f_n, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right] \\ &= (\mathbb{E} - \mathbb{E}_n)\left[\ell(\hat{f}, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right] + \mathbb{E}_n\left[\ell(f_n, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right]. \end{aligned} \tag{A.1}$$

Equation (A.1) is the main decomposition that begins the proof. The decomposition must be done this way because of the above notes regarding f_* and f_n. The first term is the empirical process term that will be treated in the subsequent subsection. For the second term in (A.1), the bias term or approximation error, we apply Bernstein's inequality to find that, with probability at least

$1 - e^{-\gamma}$,

$$\begin{aligned} \mathbb{E}_n\left[\ell(f_n, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right] &\leq \mathbb{E}\left[\ell(f_n, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z})\right] + \sqrt{\frac{2C_\ell^2 \|f_n - f_*\|_\infty^2 \gamma}{n}} + \frac{14 C_\ell M \gamma}{3n} \\ &\leq \frac{1}{c_2}\mathbb{E}\left[\|f_n - f_*\|^2\right] + \sqrt{\frac{2C_\ell^2 \|f_n - f_*\|_\infty^2 \gamma}{n}} + \frac{14 C_\ell M \gamma}{3n} \\ &\leq \frac{1}{c_2}\epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \gamma}{n}} + \frac{14 C_\ell M \gamma}{3n}, \end{aligned} \tag{A.2}$$

using Lemma 9 (wherein c_2 is given) and $\mathbb{E}\left[\|f_n - f_*\|^2\right] \leq \|f_n - f_*\|_\infty^2$, along with the definition of ϵ_n^2, and Lemma 10.

Once the empirical process term is controlled (in Section A.2), the two bounds will be brought back together to compute the final result, see Section A.3.

A.2 Localized Analysis

We now turn to bounding the first term in (A.1) (the empirical processes term) using a novel localized analysis that derives bounds based on scale insensitive complexity measure. The ideas of our localization are rooted in Koltchinskii and Panchenko (2000) and Bartlett et al. (2005). This proof section proceeds in several steps.

A key quantity is the Rademacher complexity of the function class at hand. Given i.i.d. Rademacher draws, $\eta_i = \pm 1$ with equal probability independent of the data, the random variable $R_n\mathcal{F}$, for a function class $\mathcal{F}$, is defined as

$$R_n\mathcal{F} := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \eta_i f(x_i).$$

Inuitively, $R_n\mathcal{F}$ measures how flexible the function class is for predicting random signs. Taking the expectation of $R_n\mathcal{F}$ conditioned on the data we obtain the *empirical Rademacher complexity*, denoted $\mathbb{E}_\eta[R_n\mathcal{F}]$. When the expectation is taken over both the data and the draws η_i, $\mathbb{E}R_n\mathcal{F}$, we get the *Rademacher complexity*.

A.2.1 Step I: Quadratic Process

The first step is to show that the empirical L_2 norm of $(f - f_*)$ is at most twice the population bound, for certain functions f outside a certain critical radius. This fact will be used later on. Denote $\|f\|_n := \left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2\right)^{1/2}$ to be the empirical L_2-distance. To do so, we study the quadratic process

$$\|f - f_*\|_n^2 - \|f - f_*\|_{L_2(X)}^2 = \mathbb{E}_n(f - f_*)^2 - \mathbb{E}(f - f_*)^2.$$

We will apply the symmetrization of Lemma 6 to $g = (f - f_*)^2$ restricted to a radius $\|f - f_*\|_{L_2(X)} \leq r$. This function g has variance bounded as

$$\mathbb{V}[g] \leq \mathbb{E}[g^2] \leq \mathbb{E}((f - f_*)^4) \leq 4M^2 r^2.$$

Writing $g = (f + f_*)(f - f_*)$, we see that by Assumption 1, $|g| \leq 2M|f - f_*| \leq 4M^2$, where the first inequality verifies that g has a Lipschitz constant of $2M$, and second that g itself is bounded. We therefore apply Lemma 6, to obtain, with probability at least $1 - \exp(-\gamma)$, that for any $f \in \mathcal{F}$ with $\|f - f_*\|_{L_2(X)} \leq r$,

$$\begin{aligned}
&\mathbb{E}_n(f - f_*)^2 - \mathbb{E}(f - f_*)^2 \\
&\qquad \leq 3\mathbb{E}R_n\{g = (f - f_*)^2 : f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq r\} + 2Mr\sqrt{\frac{2\gamma}{n}} + \frac{16M^2}{3}\frac{\gamma}{n} \\
&\qquad \leq 6M\mathbb{E}R_n\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq r\} + 2Mr\sqrt{\frac{2\gamma}{n}} + \frac{16M^2}{3}\frac{\gamma}{n}, \qquad \text{(A.3)}
\end{aligned}$$

where the second inequality applies Lemma 2 to the Lipschitz functions $\{g\}$.

Suppose the radius r satisfies

$$r^2 \geq 6M\mathbb{E}R_n\{f - f_* : f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq r\} \tag{A.4}$$

and

$$r^2 \geq \frac{8M^2\gamma}{n}. \tag{A.5}$$

Then we conclude from from (A.3) that

$$\mathbb{E}_n(f - f_*)^2 \leq r^2 + r^2 + 2Mr\sqrt{\frac{2\gamma}{n}} + \frac{16M^2}{3}\frac{\gamma}{n} \leq (2r)^2 \tag{A.6}$$

where the first inequality uses (A.4) and the second line uses (A.5). This means that for r above the “critical radius” (see **Step III**), the empirical L_2-norm is at most twice the population one with probability at least $1 - \exp(-\gamma)$.

A.2.2 Step II: One Step Improvement

In this step we will show that given a bound on $\|\hat{f} - f_*\|_{L_2(X)}$ we can use this bound as information to obtain a tighter bound, if the initial bound is loose as made precise at the end of this step. Suppose we know that for some r_0, $\|\hat{f} - f_*\|_{L_2(X)} \leq r_0$ and, by **Step I**, $\|\hat{f} - f_*\|_n \leq 2r_0$. We may always start with $r_0 = 2M$ given Assumption 1 and (2.3). Apply Lemma 6 with $\mathcal{G} := \{g = \ell(f, \boldsymbol{z}) - \ell(f_*, \boldsymbol{z}) : f \in \mathcal{F}_{\mathrm{DNN}}, \|f - f_*\|_{L_2(X)} \leq r_0\}$, we find that, with probability at least $1 - 2e^{-\gamma}$, the empirical process term of (A.1) is bounded as

$$(\mathbb{E}-\mathbb{E}_n)\left[\ell(\hat{f},\boldsymbol{z})-\ell(f_*,\boldsymbol{z})\right] \leq 6\mathbb{E}_\eta R_n\mathcal{G}+\sqrt{\frac{2C_\ell^2 r_0^2\gamma}{n}}+\frac{46MC_\ell}{3}\frac{\gamma}{n}, \tag{A.7}$$

where the middle term is due to the following variance calculation (recall Lemma 10)

$$\mathbb{V}[g] \leq \mathbb{E}[g^2] = \mathbb{E}[|\ell(f,\boldsymbol{z})-\ell(f_*,\boldsymbol{z})|^2] \leq C_\ell^2\mathbb{E}(f-f_*)^2 \leq C_\ell^2 r_0^2$$

and the final term follows because the right side of (C.1) is bounded by $C_\ell 2M$. Here the fact that Lemma 6 is variance dependent, and that the variance depends on the radius r_0, is important. It is this property which enables a sharpening of the rate with step-by-step reductions in the variance bound, as in Section A.2.4.

For the empirical Rademacher complexity term, the first term of (A.7), Lemma 2, **Step I**, and Lemma 3, yield

$$\begin{aligned}
\mathbb{E}_\eta R_n\mathcal{G} &= \mathbb{E}_\eta R_n\{g : g=\ell(f,\boldsymbol{z})-\ell(f_*,\boldsymbol{z}), f\in\mathcal{F}_{\mathrm{DNN}}, \|f-f_*\|\leq r_0\}\\
&\leq C_\ell\mathbb{E}_\eta R_n\{f-f_* : f\in\mathcal{F}_{\mathrm{DNN}}, \|f-f_*\|\leq r_0\}\\
&\leq C_\ell\mathbb{E}_\eta R_n\{f-f_* : f\in\mathcal{F}_{\mathrm{DNN}}, \|f-f_*\|_n\leq 2r_0\}\\
&\leq C_\ell \inf_{0<\alpha<2r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{DNN}},\|\cdot\|_n)}d\delta\right\}\\
&\leq C_\ell \inf_{0<\alpha<2r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{DNN}}|_{x_1,\ldots,x_n},\infty)}d\delta\right\}\quad ,
\end{aligned}$$

Recall Lemma 4, one can further upper bound the entropy integral when $n>\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})$,

$$\begin{aligned}
&\inf_{0<\alpha<2r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\log\mathcal{N}(\delta,\mathcal{F}_{\mathrm{DNN}}|_{x_1,\ldots,x_n},\infty)}d\delta\right\}\\
&\leq \inf_{0<\alpha<2r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})\log\frac{2eMn}{\delta\cdot\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})}}d\delta\right\}\\
&\leq 32r_0\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})}{n}}\left(\log\frac{2eM}{r_0}+\frac{3}{2}\log n\right)
\end{aligned}$$

with a particular choice of $\alpha=2r_0\sqrt{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})/n}<2r_0$. Therefore, whenever $r_0\geq 1/n$ and $n\geq(2eM)^2$,

$$\mathbb{E}_\eta R_n\mathcal{G}\leq 64C_\ell r_0\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})}{n}}\log n.$$

Applying this bound to (A.7), we have

$$(\mathbb{E}-\mathbb{E}_n)\left[\ell(\hat{f},\boldsymbol{z})-\ell(f_*,\boldsymbol{z})\right] \leq Kr_0\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})}{n}}\log n+r_0\sqrt{\frac{2C_\ell^2\gamma}{n}}+\frac{46MC_\ell}{3}\frac{\gamma}{n} \tag{A.8}$$

where $K = 6 \times 64 C_\ell$.

Going back now to the main decomposition, plug (A.8) and (A.2) into (A.1), and we find that

$$\begin{aligned}
&c_2 \|\hat{f} - f_*\|^2_{L_2(X)} \\
&\leq K r_0 \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})}{n}} \log n + r_0 \sqrt{\frac{2C_\ell^2 \gamma}{n}} + \frac{46 M C_\ell}{3} \frac{\gamma}{n} + \left(\frac{1}{c_2} \epsilon^2 + \epsilon \sqrt{\frac{2C_\ell^2 \gamma}{n}} + \frac{14 C_\ell M \gamma}{3n} \right) \\
&\leq r_0 \cdot \left(K \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{DNN}})}{n}} \log n + \sqrt{\frac{2C_\ell^2 \gamma}{n}} \right) + \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \gamma}{n}} + 20 M C_\ell \frac{\gamma}{n} \\
&\leq r_0 \cdot \left(K \sqrt{C} \sqrt{\frac{W L \log W}{n}} \log n + \sqrt{\frac{2C_\ell^2 \gamma}{n}} \right) + \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \gamma}{n}} + 20 M C_\ell \frac{\gamma}{n}, \qquad \text{(A.9)}
\end{aligned}$$

with probability at least $1 - 3\exp(-\gamma)$, where the last line applies Lemma 7. Therefore, whenever $\epsilon_n \ll r_0$ and $\sqrt{\frac{WL \log W}{n}} \log n \ll r_0$, the knowledge that $\|\hat{f} - f_*\|_{L_2(X)} \leq r_0$ implies that (with high probability) $\|\hat{f} - f_*\|_{L_2(X)} \leq r_1$, for $r_1 \ll r_0$. One can recursively improve the bound r to a fixed point/radius r_*, which describes the fundamental difficulty of the problem. This is done in the course of the next two steps.

A.2.3 Step III: Critical Radius

Formally, define the critical radius r_* to be the largest fixed point

$$r_* = \inf \left\{ r > 0 : 6 M \mathbb{E} R_n \{ f - f_* : f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq s \} < s^2, \forall u \geq r \right\}.$$

By construction this obeys (A.4), and thus so does $2r_*$. Denote the event E (depending on the data) to be

$$E = \left\{ \|f - f_*\|_n \leq 4 r_*, \quad \text{for all } f \in \mathcal{F} \text{ and } \quad \|f - f_*\|_{L_2(X)} \leq 2 r_* \right\}$$

and $\mathbb{1}_E$ to be the indicator that event E holds. We know from (A.6) that $\mathbb{P}(\mathbb{1}_E = 1) \geq 1 - n^{-1}$, provided $r_* \geq \sqrt{8} M \sqrt{\log n / n}$ to satisfy (A.5).

We can now give an upper bound for the the critical radius r_*. Using the logic of **Step II** to bound the empirical Rademacher complexity, and then applying Lemma 7, we find that

$$\begin{aligned}
r_*^2 &\leq 6 M \mathbb{E} R_n \left\{ f - f_* : f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq r_* \right\} \\
&\leq 6 M \mathbb{E} R_n \left\{ f - f_* : f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq 2 r_* \right\} \\
&\leq 6 M \mathbb{E} \left\{ \mathbb{E}_\eta R_n \{ f - f_* : f \in \mathcal{F}, \|f - f_*\|_n \leq 4 r_* \} \mathbb{1}_E + 2M (1 - \mathbb{1}_E) \right\} \\
&\leq 12 M K \sqrt{C} \cdot r_* \sqrt{\frac{W L \log W}{n}} \log n + 12 M^2 \frac{1}{n}
\end{aligned}$$

$$\leq 14MK\sqrt{C} \cdot r_* \sqrt{\frac{WL \log W}{n} \log n},$$

with the last line relying on the above restriction that $r_* \geq \sqrt{8}M\sqrt{\log n/n}$. Dividing through by r_* yields the final bound:

$$r_* \leq 14MK\sqrt{C}\sqrt{\frac{WL \log W}{n} \log n}. \tag{A.10}$$

A.2.4 Step III: Localization

Divide the space $\mathcal{F}_{\mathrm{DNN}}$ into shells of increasing radius by intersecting it with the balls

$$B(f_*, \bar{r}), B(f_*, 2\bar{r})\backslash B(f_*, \bar{r}), \ldots B(f_*, 2^l\bar{r})\backslash B(f_*, 2^{l-1}\bar{r}) \tag{A.11}$$

where $l \leq \log_2 \frac{2M}{\sqrt{(\log n)/n}}$. We will specify the choice of $\bar{r}$ shortly.

Suppose $\bar{r} > r_*$. Then for each shell, **Step I** implies that with probability at least $1 - 2l\exp(-\gamma)$,

$$\|f - f_*\|_{L_2(X)} \leq 2^j\bar{r} \;\Rightarrow\; \|f - f_*\|_n \leq 2^{j+1}\bar{r}. \tag{A.12}$$

Further, suppose that for some $j \leq l$

$$\hat{f} \in B(f_*, 2^j\bar{r})\backslash B(f_*, 2^{j-1}\bar{r}). \tag{A.13}$$

Then applying the one step improvement argument in **Step II** (again the variance dependence captured in Lemma 6 is crucial, here reflected in the variance within each shell), Equation (A.9) yields that with probability at least $1 - 3\exp(-\gamma)$,

$$\begin{aligned}\|\hat{f} - f_*\|^2_{L_2(X)} &\leq \frac{1}{c_2}\left\{2^j\bar{r} \cdot \left(K\sqrt{C}\sqrt{\frac{WL\log W}{n}\log n} + \sqrt{\frac{2C_\ell^2 t}{n}}\right) + \epsilon_n^2 + \epsilon_n\sqrt{\frac{2C_\ell^2\gamma}{n}} + 20MC_\ell\frac{\gamma}{n}\right\} \\ &\leq 2^{2j-2}\bar{r}^2,\end{aligned}$$

if the following two conditions hold:

$$\frac{1}{c_2}\left(K\sqrt{C}\sqrt{\frac{WL\log W}{n}\log n} + \sqrt{\frac{2C_\ell^2\gamma}{n}}\right) \leq \frac{1}{8}2^j\bar{r}$$

$$\frac{1}{c_2}\left(\epsilon_n^2 + \epsilon_n\sqrt{\frac{2C_\ell^2 t}{n}} + 26MC_\ell\frac{\gamma}{n}\right) \leq \frac{1}{8}2^{2j}\bar{r}^2.$$

It is easy to see that these two hold for all j if we choose

$$\bar{r} = \frac{8}{c_2}\left(K\sqrt{C}\sqrt{\frac{WL\log W}{n}}\log n + \sqrt{\frac{2C_\ell^2\gamma}{n}}\right) + \left(\sqrt{\frac{16}{c_2}}\epsilon_n + \sqrt{\frac{208MC_\ell}{c_2}\frac{\gamma}{n}}\right) + r_*. \tag{A.14}$$

Therefore with probability at least $1-5l\exp(-\gamma)$, we can perform shell-by-shell argument combining the results in **Step I** and **Step II**:

$$\begin{aligned}
&\|\hat{f}-f_*\|_{L_2(X)} \le 2^l\bar{r} \ \text{ and } \ \|\hat{f}-f_*\|_n \le 2^{l+1}\bar{r} \\
&\qquad \text{implies } \ \|\hat{f}-f_*\|_{L_2(X)} \le 2^{l-1}\bar{r} \ \text{ and } \ \|\hat{f}-f_*\|_n \le 2^{l}\bar{r} \\
&\qquad\quad \cdots\cdots \\
&\qquad\qquad \text{implies } \ \|\hat{f}-f_*\|_{L_2(X)} \le 2^{0}\bar{r} \ \text{ and } \ \|\hat{f}-f_*\|_n \le 2^{1}\bar{r}.
\end{aligned}$$

The "and" part of each line follows from **Step I** and the implication uses the above argument following **Step II**. Therefore in the end, we conclude with probability at least $1-5l\exp(-t)$,

$$\|\hat{f}-f_*\|_{L_2(X)} \le \bar{r} \ , \tag{A.15}$$

$$\|\hat{f}-f_*\|_n \le 2\bar{r} \ . \tag{A.16}$$

Therefore choose $\gamma = \log(5l)+\gamma'$, we know from (A.14), and the upper bound on r_* in (A.10)

$$\begin{aligned}
\bar{r} &\le \frac{8}{c_2}\left(K\sqrt{C}\sqrt{\frac{WL\log W}{n}}\log n + \sqrt{\frac{2C_\ell^2(\log\log n+\gamma')}{n}}\right) + \left(\sqrt{\frac{16}{c_2}}\epsilon_n + \sqrt{\frac{208MC_\ell}{c_2}\frac{\log\log n+\gamma'}{n}}\right) + r_* \\
&\le C'\left(\sqrt{\frac{WL\log W}{n}}\log n + \sqrt{\frac{\log\log n+\gamma'}{n}} + \epsilon_n\right),
\end{aligned} \tag{A.17}$$

with some absolute constant $C'>0$. This completes the proof of Theorem 2.

A.3 Final Steps for the MLP case

For the multi-layer perceptron, $W \le C\cdot H^2L$, and plugging this into the bound (A.17), we obtain

$$C'\left(\sqrt{\frac{H_n^2L_n^2\log(H_n^2L_n)}{n}}\log n + \sqrt{\frac{\log\log n+t'}{n}} + \epsilon_n\right)$$

To optimize this upper bound on $\bar{r}$, we need to specify the trade-offs in ϵ_n and H_n and L_n. To do so, we utilize the MLP-specific approximation rate of Lemma 8 and the embedding of Lemma 1. Lemma 1 implies that, for any ϵ_n, one can embed the approximation class $\mathcal{F}_{\text{DNN}}$ given by Lemma

8 into a standard MLP architecture $\mathcal{F}_{\mathrm{MLP}}$, where specifically

$$\begin{aligned} H_n &= H(\epsilon_n) \le W(\epsilon_n)L(\epsilon_n) \le C^2 \epsilon_n^{-\frac{d}{\beta}} (\log(1/\epsilon_n)+1)^2, \\ L_n &= L(\epsilon_n) \le C \cdot (\log(1/\epsilon_n)+1). \end{aligned}$$

For standard MLP architecture $\mathcal{F}_{\mathrm{MLP}}$,

$$H_n^2 L_n^2 \log(H_n^2 L_n) \le \tilde{C} \cdot \epsilon_n^{-\frac{2d}{\beta}} (\log(1/\epsilon_n)+1)^7.$$

Thus we can optimize the upper bound

$$\bar{r} \le C' \left(\sqrt{\frac{\epsilon_n^{-\frac{2d}{\beta}} (\log(1/\epsilon_n)+1)^7}{n} \log n} + \sqrt{\frac{\log\log n + \gamma'}{n}} + \epsilon_n \right)$$

by choosing $\epsilon_n = n^{-\frac{\beta}{2(\beta+d)}}$, $H_n = c_1 \cdot n^{\frac{d}{2(\beta+d)}} \log^2 n$, $L_n = c_2 \cdot \log n$. This gives

$$\bar{r} \le C \left(n^{-\frac{\beta}{2(\beta+d)}} \log^4 n + \sqrt{\frac{\log\log n + t'}{n}} \right).$$

Hence putting everything together, with probability at least $1 - \exp(-\gamma)$,

$$\begin{aligned} \mathbb{E}(\hat{f} - f_*)^2 &\le \bar{r}^2 \le C \left(n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n + \gamma}{n} \right), \\ \mathbb{E}_n(\hat{f} - f_*)^2 &\le (2\bar{r})^2 \le 4C \left(n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n + \gamma}{n} \right). \end{aligned}$$

This completes the proof of Theorem 1.

B Proof of Corollaries 1 and 2

For Corollary 1, we want to optimize

$$\frac{WL\log U}{n} \log n + \frac{\log\log n + \gamma}{n} + \epsilon_{\mathrm{DNN}}^2.$$

Yarotsky (2017, Theorem 1) shows that for the approximation error ϵ_{DNN} to obey $\epsilon_{\mathrm{DNN}} \le \epsilon$, it suffices to choose $W, U \propto \epsilon^{-\frac{d}{\beta}}(\log(1/\epsilon)+1)$ and $L \propto (\log(1/\epsilon)+1)$, given the specific architecture described therein. Therefore, we attain $\epsilon \asymp n^{-\beta/(2\beta+d)}$ by setting $W, U \asymp n^{d/(2\beta+d)}$ and $L \asymp \log n$, yielding the desired result.

For Corollary 2, we need to optimize

$$\frac{H^2 L_2 \log(HL)}{n} \log n + \frac{\log\log n + \gamma}{n} + \epsilon_{\text{MLP}}^2.$$

Yarotsky (2018, Theorem 1) shows that for the approximation error ϵ_{MLP} to obey $\epsilon_{\text{MLP}} \leq \epsilon$, it suffices to choose $H \propto 2d + 10$ and $L \propto \epsilon^{-\frac{d}{2}}$, given the specific architecture described therein. Thus, for $\epsilon \asymp n^{-1/(2+d)}$ we take $L \asymp n^{-d/(4+2d)}$, and the result follows.

C Supporting Lemmas

First, we show that one can embed a feedforward network into the multi-layer perceptron architecture by adding auxiliary hidden nodes. This idea is due to Yarotsky (2018).

Lemma 1 (Embedding). *For any function $f \in \mathcal{F}_{\text{DNN}}$, there is a $g \in \mathcal{F}_{\text{MLP}}$, with $H \leq WL + U$, such that $g = f$.*

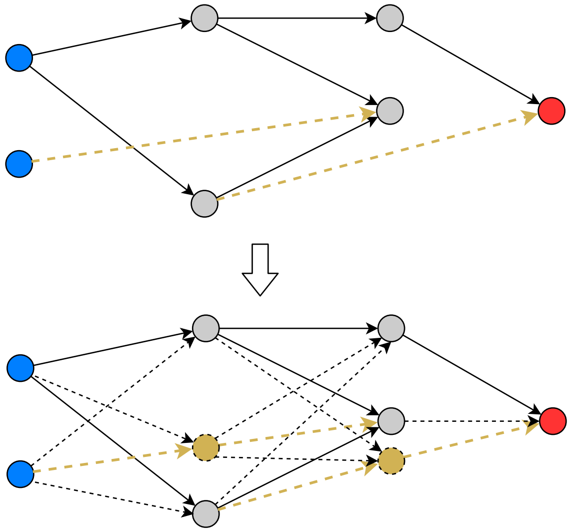


Figure 6: Illustration of how to embed a feedforward network into a multi-layer perceptron, with auxiliary hidden nodes (shown in yellow).

Proof. The idea is illustrated in Figure 6. For the edges in the directed graph of $f \in \mathcal{F}_{\text{DNN}}$ that connect nodes not in adjacent layers (shown in yellow in Figure 6), one can insert auxiliary hidden units in order to simply "pass forward" the information. The number of such auxiliary "passforward units" is at most the number of offending edges times the depth L (i.e. for each edge, at most L auxiliary nodes are required), and this is bounded by WL. Therefore the width of the MLP network that subsumes the original is upper bounded by $WL + U$ while still maintaining the required embedding that for any $f_\theta \in \mathcal{F}_{\text{DNN}}$, there is a $g_{\theta'} \in \mathcal{F}_{\text{MLP}}$ such that $g_{\theta'} = f_\theta$. In order to match modern practice we only need to show that auxiliary units can be implemented with ReLU activation. This can be done by setting the constant ("bias") term b of each auxiliary unit large enough to ensure $\sigma(\tilde{\boldsymbol{x}}'\boldsymbol{w} + b) = \tilde{\boldsymbol{x}}'\boldsymbol{w} + b$, and then subtracting the same b in the last receiving unit along the path. □

Next, we give two properties of the Rademacher complexity that we require (see Mendelson, 2003).

Lemma 2 (Contraction)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz contraction $|\phi(x) - \phi(y)| \leq L|x - y|$, then*

$$\mathbb{E}_\eta R_n \phi\{\phi \circ f : f \in \mathcal{F}\} \leq L\mathbb{E}_\eta R_n \mathcal{F}.$$

Lemma 3 (Dudley's Chaining)**.** *Let $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_n)$ denote the metric entropy for class $\mathcal{F}$ (with covering radius δ and metric $\|\cdot\|_n$), then*

$$\mathbb{E}_\eta R_n\{f : f \in \mathcal{F}, \|f\|_n \leq r\} \leq \inf_{0<\alpha<r} \left\{4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^r \sqrt{\log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_n)} d\delta \right\} .$$

Furthermore, because $\|f\|_n \leq \max_i |f(\boldsymbol{x}_i)|$, and therefore $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_n) \leq \mathcal{N}(\delta, \mathcal{F}|_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}, \infty)$ and so the upper bound in the conclusions also holds with $\mathcal{N}(\delta, \mathcal{F}|_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}, \infty)$.

The next two results, Theorems 12.2 and 14.1 in Anthony and Bartlett (1999), show that the metric entropy may be bounded in terms of the pseudo-dimension and that the latter is bounded by the VapnikChervonenkis (VC) dimension.

Lemma 4. *Assume for all $f \in \mathcal{F}$, $\|f\| \leq M$. Denote the pseudo-dimension of $\mathcal{F}$ as $\mathrm{Pdim}(\mathcal{F})$, then for $n \geq \mathrm{Pdim}(\mathcal{F})$, we have for any δ,*

$$\mathcal{N}(\delta, \mathcal{F}|_{x_1,\ldots,x_n}, \infty) \leq \left(\frac{2eM \cdot n}{\delta \cdot \mathrm{Pdim}(\mathcal{F})}\right)^{\mathrm{Pdim}(\mathcal{F})} .$$

Lemma 5. *If $\mathcal{F}$ is the class of functions generated by a neural network with a fixed architecture and fixed activation functions, then*

$$\mathrm{Pdim}(\mathcal{F}) \leq \mathrm{VCdim}(\tilde{\mathcal{F}})$$

where $\tilde{\mathcal{F}}$ has only one extra input unit and one extra computation unit compared to $\mathcal{F}$.

The following symmetrization lemma bounds the empirical processes term using Rademacher complexity, and is thus a crucial piece of our localization. This is a standard result based on Talagrand's concentration, but here special care is taken with the dependence on the variance.

Lemma 6 (Symmetrization, Theorem 2.1 in Bartlett et al. (2005))**.** *For any $g \in \mathcal{G}$, assume that $|g| \leq G$ and $\mathbb{V}[g] \leq V$. Then for every $\gamma > 0$, with probability at least $1 - e^{-\gamma}$*

$$\sup_{g \in \mathcal{G}} \{\mathbb{E}g - \mathbb{E}_n g\} \leq 3\mathbb{E}R_n\mathcal{G} + \sqrt{\frac{2V\gamma}{n}} + \frac{4G}{3}\frac{\gamma}{n} ,$$

and with probability at least $1-2e^{-t}$

$$\sup_{g\in\mathcal{G}}\{\mathbb{E}g-\mathbb{E}_n g\}\leq 6\mathbb{E}_\eta R_n\mathcal{G}+\sqrt{\frac{2V\gamma}{n}}+\frac{23G}{3}\frac{\gamma}{n}\ .$$

The same result holds for $\sup_{g\in\mathcal{G}}\{\mathbb{E}_n g-\mathbb{E}g\}$.

In turn, when bounding the complexity using the VC dimension of $\mathcal{F}_{\mathrm{DNN}}$, we use the following bounds on the latter.

Lemma 7 (Theorem 6 in Bartlett et al. (2017), ReLU case)**.** *Consider a ReLU network architecture* $\mathcal{F}=\mathcal{F}_{\mathrm{DNN}}(W,L,U)$*, then the VC-dimension and pseudo-dimension is sandwiched by*

$$c\cdot WL\log(W/L)\leq \mathrm{VCdim}(\mathcal{F})\leq C\cdot WL\log W,$$

with some universal constants $c,C>0$*. The same result holds for* $\mathrm{Pdim}(\mathcal{F})$*.*

For multi-layer perceptrons we use the following approximation result, Theorem 1 of Yarotsky (2017).

Lemma 8. *There exists a network class* $\mathcal{F}_{\mathrm{DNN}}$*, with ReLU activation, such that for any* $\epsilon>0$*:*

(a) $\mathcal{F}_{\mathrm{DNN}}$ *approximates the* $W^{\beta,\infty}([-1,1]^d)$ *in the sense for any* $f_*\in W^{\beta,\infty}([-1,1]^d)$*, there exists a* $f_n(\epsilon):=f_n\in\mathcal{F}_{\mathrm{DNN}}$ *such that*

$$\|f_n-f_*\|_\infty\leq\epsilon,$$

(b) *and* $\mathcal{F}_{\mathrm{DNN}}$ *has* $L(\epsilon)\leq C\cdot(\log(1/\epsilon)+1)$ *and* $W(\epsilon),U(\epsilon)\leq C\cdot\epsilon^{-\frac{d}{\beta}}(\log(1/\epsilon)+1)$*.*

Here C *only depends on* d *and* β*.*

The next two lemmas relate differences in the loss to differences in the functions themselves. These are used at very steps and follow from standard calculus arguments.

Lemma 9 (Curvature)**.** *For Both the least squares* (2.1) *and logistic* (2.2) *loss functions*

$$c_1\mathbb{E}\left[(f-f_*)^2\right]\leq\mathbb{E}[\ell(f,\boldsymbol{Z})]-\mathbb{E}\ell[(f_*,\boldsymbol{Z})]\leq c_2\mathbb{E}\left[(f-f_*)^2\right].$$

For least squares, $c_1=c_2=1/2$*. For logistic,* $c_1=(2(\exp(M)+\exp(-M)+2))^{-1}$*,* $c_2=1/8$*.*

Proof. For least squares, using iterated expectations

$$\begin{aligned}2\mathbb{E}\ell(f,\boldsymbol{Z})-2\mathbb{E}\ell(f_*,\boldsymbol{Z})&=\mathbb{E}\left[-2Yf+f^2+2Yf_*-f_*^2\right]\\&=\mathbb{E}\left[-2f_*f(\boldsymbol{x})+f^2+2(f_*)^2-f_*^2\right]\\&=\mathbb{E}\left[(f-f_*)^2\right].\end{aligned}$$

For logistic regression,

$$\mathbb{E}[\ell(f,\boldsymbol{Z})]-\mathbb{E}[\ell(f_*,\boldsymbol{Z})]=\mathbb{E}\left[-\frac{\exp(f_*)}{1+\exp(f_*)}(f-f_*)+\log\left(\frac{1+\exp(f)}{1+\exp(f_*)}\right)\right].$$

Define $h_a(b)=-\frac{\exp(a)}{1+\exp(a)}(b-a)+\log\left(\frac{1+\exp(b)}{1+\exp(a)}\right)$, then

$$h_a(b)=h_a(a)+h_a'(a)(b-a)+\frac{1}{2}h_a''\left(\xi a+(1-\xi)b\right)(b-a)^2$$

and $h_a''(b)=\frac{1}{\exp(b)+\exp(-b)+2}\leq\frac{1}{4}$. The lower bound holds as $|\xi f_*+(1-\xi)f|\leq M$. □

Lemma 10 (Lipschitz)**.** *For Both the least squares* (2.1) *and logistic* (2.2) *loss functions*

$$|\ell(f,\boldsymbol{z})-\ell(g,\boldsymbol{z})|\leq C_\ell|f(\boldsymbol{x})-g(\boldsymbol{x})|. \tag{C.1}$$

For least squares, $C_\ell=M$. For logistic regression, $C_\ell=1$.

Our last result is to verify condition (c) of Theorem 3. We do so using our localization, which may be of future interest in second-step inference with machine learning methods.

Lemma 11. *Let the conditions of Theorem 3 hold. Then*

$$\mathbb{E}_n\left[(\hat\mu_t(\boldsymbol{x}_i)-\mu_t(\boldsymbol{x}_i))\left(1-\frac{\mathbb{1}\{t_i=t\}}{\mathbb{P}[T=t|\boldsymbol{X}=\boldsymbol{x}_i]}\right)\right]=o_P\left(n^{-\frac{\beta}{\beta+d}}\log^8 n+\frac{\log\log n}{n}\right)=o_P\left(n^{-1/2}\right).$$

Proof. Without loss of generality we can take $\bar p<1/2$. The only estimated function here is $\mu_t(\boldsymbol{x})$, which plays the role of f_* here. For function(als) $L(\cdot)$ of the form

$$L(f):=(f(\boldsymbol{x}_i)-f_*(\boldsymbol{x}_i))\left(1-\frac{\mathbb{1}\{t_i=t\}}{\mathbb{P}[T=t|\boldsymbol{X}=\boldsymbol{x}_i]}\right),$$

it is true that

$$\mathbb{E}[L(f)]=\mathbb{E}\left[(f(\boldsymbol{X})-f_*(\boldsymbol{X}))\left(1-\frac{\mathbb{E}[\mathbb{1}\{t_i=t\}|\boldsymbol{x}_i]}{\mathbb{P}[T=t|\boldsymbol{X}=\boldsymbol{x}_i]}\right)\right]=0$$

and

$$\begin{aligned}\mathbb{V}[L(f)]&\leq(1/\bar p-1)^2\,\mathbb{E}\left[(f(\boldsymbol{X})-f_*(\boldsymbol{X}))^2\right]\leq(1/\bar p-1)^2\,\bar r^2\\|L(f)|&\leq(1/\bar p-1)\,2M.\end{aligned}$$

For $\bar r$ defined in (A.14),

$$\begin{aligned}&6M\mathbb{E}R_n\{f-f_*:f\in\mathcal{F},\|f-f_*\|_{L_2(X)}\leq\bar r\}\leq\bar r^2\\&\qquad\mathbb{E}R_n\{L(f):f\in\mathcal{F},\|f-f_*\|_{L_2(X)}\leq\bar r\}\leq(1/\bar p-1)\,\mathbb{E}R_n\{f-f_*:f\in\mathcal{F},\|f-f_*\|_{L_2(X)}\leq\bar r\}\end{aligned}$$

where the first line is due to $\bar{r} > r_*$, and second line uses Lemma 2.

Then by the localization analysis and Lemma 6, for all $f \in \mathcal{F}, \|f - f_*\|_{L_2(X)} \leq \bar{r}$, $L(f)$ obeys

$$\begin{aligned}\mathbb{E}_n[L(f)] = \mathbb{E}_n[L(f)] - \mathbb{E}[L(f)] &\leq 3C\bar{r}^2 + \bar{r}\sqrt{\frac{2\left(1/\bar{p}-1\right)^2 t}{n}} + \frac{4\left(1/\bar{p}-1\right)2M}{3}\frac{t}{n} \leq 4C\bar{r}^2 \\ &\leq C \cdot \left\{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\right\}, \\ \sup_{f \in \mathcal{F}, \|f-f_*\|_{L_2(X)} \leq \bar{r}} \mathbb{E}_n[L(f)] &\leq C \cdot \left\{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\right\}.\end{aligned}$$

With probability at least $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$, $\hat{f}_{\mathrm{MLP}}$ lies in this set of functions, and therefore

$$\mathbb{E}_n[L(\hat{f}_{\mathrm{MLP}})] = \mathbb{E}_n\left[(\widehat{f}_{n,H,L}(x) - f_*(x))\left(1 - \frac{1(T=t)}{P(T=t|\boldsymbol{x}=x)}\right)\right] \leq C \cdot \left\{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\right\},$$

as claimed.

□