DEEP NEURAL NETWORKS FOR NONPARAMETRIC INTERACTION MODELS WITH DIVERGING DIMENSION

BY SOHOM BHATTACHARYA[1,a], JIANQING FAN[1,a] AND DEBARGHYA MUKHERJEE[1,a]

[1]*Department of Operations Research and Financial Engineering, Princeton University,* [a]

Deep neural networks have achieved tremendous success due to their representation power and adaptation to low-dimensional structures. Their potential for estimating structured regression functions has been recently established in the literature. However, most of the studies require the input dimension to be fixed and consequently ignore the effect of dimension on the rate of convergence and hamper their applications to modern big data with high dimensionality. In this paper, we bridge this gap by analyzing a k^{th} order nonparametric interaction model in both growing dimension scenarios (d grows with n but at a slower rate) and in high dimension ($d \gtrsim n$). In the latter case, sparsity assumptions and associated regularization are required in order to obtain optimal rates of convergence. A new challenge in diverging dimension setting is in calculation mean-square error, the covariance terms among estimated additive components are an order of magnitude larger than those of the variances and they can deteriorate statistical properties without proper care. We introduce a critical debiasing technique to amend the problem. We show that under certain standard assumptions, debiased deep neural networks achieve a minimax optimal rate both in terms of (n, d). Our proof techniques rely crucially on a novel debiasing technique that makes the covariances of additive components negligible in the mean-square error calculation. In addition, we establish the matching lower bounds.

1. Introduction. Recent advances in technology have allowed statisticians to collect data on a large number of explanatory variables to predict outcomes of interest (Goodfellow et al., 2016; Fan et al., 2020). Often, the relationship between these outcomes and predictors are highly nonlinear (e.g., image data like MNIST (LeCun, 1998), CIFAR (Krizhevsky et al., 2009) etc.), yielding a practical need to investigate multivariate nonparametric regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where X_i's are explanatory variables, Y_i's are response variables and ε_i's are unobserved errors. The statistical problem here is to recover f under some minimal smoothness assumptions. A classical result of Charles Stone (Stone, 1982) shows if f is a d-variate (β, C) smooth function (precise definition given later), the minimax optimal rate of estimation is of the order $n^{-\frac{2\beta}{2\beta+d}}$, which is referred to as "curse of dimensionality", i.e., large sample size is necessary to estimate the regression function well. In particular, when β is fixed, it is easily to see that when $d \asymp \log n$, minimax consistent estimators can not be obtained.

To tackle aforementioned issue, one needs to impose a low-dimensional structure on the function f (Kpotufe and Dasgupta, 2012; Yang and Dunson, 2016; Yang and Tokdar, 2015), such as additive regression model (Stone, 1985), projection pursuit regression model (Friedman and Stuetzle, 1981), higher order interaction model (Stone, 1994), generalized higher-order interaction model (Horowitz and Mammen, 2007; Kohler and Krzyżak, 2016), single

Keywords and phrases: Deep neural networks, High dimensional statistics, Non-parametric interaction model, Minimax rate, Sparse nonparametric components.

and multi-index models (Hardle et al., 1993), etc.; see Fan and Gijbels (1996) for an overview. The key idea behind all these simplifications is to reduce the complexity of the underlying function class and mollify the effect of dimension in the rate of convergence. For example, Stone (1985) proved that for additive regression model (when $f(x) = \sum_{j=1}^{d} f_j(x_j)$) with β-smooth univariate components, the minimax optimal rate of estimation of f in terms of squared $L_2(P_X)$ loss is $C_d n^{-\frac{2\beta}{2\beta+1}}$. Fan et al. (1998) demonstrated further the adaptivity of such an additive structure and Horowitz and Mammen (2007) studied further a general class of nonparametric regression models with unknown link functions. As each component function f_j is univariate and can be estimated at a rate $n^{-\frac{2\beta}{2\beta+1}}$, the effect of dimension appears only through the multiplicative constant C_d, not in the power of n. Later, Stone (1994) extended this result for higher-order interaction model (i.e. when $f(x) = \sum_{J \in S} f_J(x_J)$, S is a collection of subsets of $\{1, \ldots, d\}$ and $|J| \leq d^\star$ for all $J \in S$) and showed that the minimax optimal rate of estimation is $C_d n^{-\frac{2\beta}{2\beta+d^\star}}$, where, as before, the effect of d appears through a multiplicative constant. Again, all of the above results assume that d is finite, which is inappropriate in many modern data science applications.

In practice, it is not enough just to obtain minimax optimal rates, efficient and easily computable estimators are warranted. Several methods like kernel, spline, and wavelet based estimators (see Section 2 of Fan and Gijbels (1996) for an overview), backfitting algorithm (Buja et al., 1989), off-the-shelf nonparametric methods like boosting (Freund et al., 1996), random forest (Breiman, 2001) have been popularly used. Recently, neural networks have emerged as an imperative tool for analyzing nonlinear relations between covariates and response variables. Deep (or multilayer) neural networks have been the backbone of the incredible advances in machine learning that have resulted in massive success in reinforcement learning, robotics, computer vision, natural language processing, and other statistical prediction tasks (Goodfellow et al., 2016). Leveraging the availability of large amounts of digitized data, deep learning has enjoyed a plethora of empirical successes. However, most of the successes are purely human engineered (i.e. achieved after tuning lots of hyperparameters), lacking theoretical guarantees. As a consequence, researchers are naturally interested in understanding the statistical properties of such useful and practical estimators. In Kohler and Krzyżak (2005), the authors established that for the standard nonparametric models as well as for the higher-order interaction models, deep neural network (henceforth DNN) based estimators are minimax rate optimal up to logarithmic factor. Later, the results are extended for a more general class of functions (namely generalized hierarchical interaction model, see Bauer and Kohler (2019)) in a series of work byKohler and Krzyżak (2016); Bauer and Kohler (2019); Schmidt-Hieber (2020); Kohler and Langer (2021); Fan et al. (2022).

The previous research in understanding the rate of convergence of DNN-based estimators assumes the underlying dimension of covariate d is fixed and consequently ignores the effect of d in the rate of the estimator. However, in the era of big data, the dimension of underlying covariates is quite large in many practical problems, often larger than the underlying sample size n. For example, in a simple image classification problem, a 28×28 image lies in $\mathbb{R}^{784}$. Similarly, in genome-wide association studies, the number of single nucleotide polymorphism (SNP) can be significantly higher than the number of individuals(Novembre et al., 2008). In that paper, the authors have analyzed a dataset consisting of 1387 individuals and $\sim 2 \times 10^5$ SNPs. In such examples, a sharper analysis quantifying the optimal dependence on d is often necessary.

In this paper, we aim to bridge this gap by analyzing the nonparametric regression model both in growing dimension (when d grows with n but $d = o(n)$) and high dimension (when $d \gtrsim n$) setup. The distinction between these two cases lies in whether or not the sparsity and regularization are required for consistent estimation, in an analogue way to the linear

model with growing dimension and sparse linear model in high dimension. We analyze the following k-way interaction model (Stone, 1994):

$$Y_i = \sum_{l=1}^{k} \sum_{(j_1<\cdots<j_l)\in[d]^l} f_{j_1,j_2,\ldots,j_l}(X_{i,\{j_1,j_2,\ldots,j_l\}}) + \epsilon_i\,, \tag{1.2}$$

where $[d]^l$ is the collection of all ordered subsequence of length l of $\{1,\ldots,d\}$. As mentioned before, $k = o(\log n)$ is necessary in order to have a consistent estimator. Hence, we will take a finite k. Such models include the well-studied additive model (Stone, 1985), namely, $f(x) = \sum_{j=1}^{d} f_j(x_j)$ and interaction models (Stone, 1994). When $d \gtrsim n$, it is necessary to impose sparsity assumption even in the high-dimensional linear model. We, therefore, consider a *sparse k-way interaction model* as follows:

$$Y_i = \sum_{l=1}^{k} \sum_{(j_1,j_2,\ldots,j_l)\in S_l} f_{j_1,j_2,\ldots,j_l}(X_{i,\{j_1,j_2,\ldots,j_l\}}) + \epsilon_i\,, \tag{1.3}$$

where $S_l \subset [d]^l$ with $|S_l| \ll d^l$. When each univariate function is constrained to be linear, the model (1.3) reduces to a sparse parametric interaction model. When $k = 1$, the model (1.3) becomes a *sparse additive model* which has been well-studied in literature (Lin and Zhang, 2006; Koltchinskii and Yuan, 2010; Ravikumar et al., 2009; Tan and Zhang, 2019; Yuan and Zhou, 2016), building upon the recent developments on penalized linear regression. For example, Tan and Zhang (2019) proposes to doubly penalize each component f_j by its empirical norm and functional semi-norm to estimate the regression function.

When $d \ll n$, we estimate the mean function of (1.2) via minimizing squared error loss (details can be found in Section 2). However, when $d \gtrsim n$, we need to penalize explicitly to enforce sparsity. In previous works on the sparse additive models, researchers typically impose two penalties: i) one to enforce sparsity (e.g. via the summation of $L_2(\mathbb{P}_n)$ norm on the component functions) and ii) another to control the complexity of the component functions (e.g., penalize with respect to RKHS norm if the component functions lie in a RKHS). In reality, implementing such a doubly penalized estimator is often computationally challenging. To overcome such difficulty, we introduce a two-step hard thresholding-based estimator, which is motivated by the seminal work of SURE independent screening (Fan and Lv, 2008; Fan and Song, 2010) and least square estimation after model selection in high dimensional linear regression model (Belloni and Chernozhukov, 2013). The idea can be briefly described as follows: first, we obtain an initial estimator (say $\hat{f}^{\text{init}}$) of the mean function by only penalizing the summation of empirical $L_2(\mathbb{P}_n)$ norm of the component functions (which can be implemented by penalizing the last layer of neural network). Analogously, this can also be thought of as a version of a group lasso penalty, where the neural network corresponding to each component is a group of variables and we penalize the sum of the norm of each group (here we take the $L_2(\mathbb{P}_n)$) norm) to enforce a component-wise sparsity. Next, we further perform a hard thresholding on the non-zero component of $\hat{f}^{\text{init}}$ to weed out the *small* non-zero components and estimate the active sets. Finally, we perform an empirical risk minimization by minimizing the squared error loss over only the components selected in the previous step to obtain the final estimator. We prove that under some fairly standard assumptions, the estimator is minimax rate optimal. We now summarize our contribution as follows.

Contribution: Our main contribution is a rigorous theoretical analysis of k-way interaction model (also known as nonparametric ANOVA model) in both growing and high dimensional setup. To achieve this goal, we also prove several results that may be of independent interest. We summarise the key contributions below:

- We analyze the nonparametric k-way interaction model using neural network based estimator both when $d = \dim(X)$ increases with n and $d = o(n)$, and when $d \gtrsim n$ with aforementioned regularization. We show that the neural network-based estimator is minimax rate optimal up to a poly-log factor.
- We introduce a novel debiasing technique that makes the covariances among additive components negligible, which reduces statistical errors.
- We prove an approximation result of smooth function using deep neural network (Theorem 2.7) via a *novel debiasing technique*, which implies that one can approximate a smooth function using neural networks at the same rate even under constraints (i.e. marginals of a multivariate function are 0).
- We establish the minimax lower bound for estimating the high dimensional k-way interaction model, which, to the best of our knowledge, is not present in the literature.

The rest of the paper is organized as follows: In Section 2, we analyze the k way interaction model when $d \ll n$. We divide the entire analysis into three subsections: Subsection 2.1 bounds approximation error, Subsection 2.2 analyzes the statistical error, and Subsection 2.3 develops the minimax lower bound. Section 3 deals with the analysis when $d \gtrsim n$. We broadly divide our analysis into two parts; Subsection 3.1 contains the analysis for fixed designs and Subsection 3.2 deals with the case of random designs. Furthermore, in Subsection 3.3, we establish the minimax lower bound for the sparse nonparametric interaction models.Section 4 provides conclusional remarks. Finally, Section 5 contains the proof of Theorem 2.7. The remaining proofs can be found in the Appendix.

2. Analysis of DNN in low diverging dimension. In this section, we present our analysis of the DNN-based estimator of the mean function when the dimension d of X grows with n but $d \ll n$. We consider a k-order interaction nonparametric regression model. For an input-output pair (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, a k-order interaction model is defined as:

$$Y_i = \sum_{\substack{J \subset [d] \\ |J| \le k}} f_{0,J}(X_{i,J}) + \epsilon_i \,. \tag{2.1}$$

Here, ϵ_i's are assumed to be centered error independent of X and the functions $f_{0,J} : \mathbb{R}^{|J|} \mapsto \mathbb{R}$. For simplicity of exposition, we henceforth confine ourselves to a 2-order interaction model. The extension of our analysis from a 2-order interaction model to a general k-order interaction model is purely technical and the proof ideas will be outlined at the end of the relevant sections. For $k = 2$, the the model presented in equation (2.1) simplifies to:

$$Y_i = \sum_{j=1}^{d} f_{0,j}(X_{i,j}) + \sum_{j<k} f_{0,jk}(X_{i,j}, X_{i,k}) + \epsilon_i \equiv f_0(X_i) + \varepsilon_i \,. \tag{2.2}$$

where $f'_{0,j}s$ are univariate functions and $f_{0,jk}$'s are bivariate functions. We will make the following assumption on X and ε:

ASSUMPTION 2.1. *X is supported on $[0,1]^d$ and admits a density function p such that $\sup_{x \in [0,1]^d} p(x) =: p_{\max} < \infty$, where $p_{\max}$ is free of d. We assume ϵ is sub-gaussian with sub-gaussian constant σ_ϵ^2.*

Assumption 2.1 is a standard assumption in the literature of nonparametric regression. Note that assuming X has compact support is equivalent to assuming that the support is $[0,1]^d$ via centering and scaling. Many of our results can be extended to the case of unbounded

X via truncation arguments – we omit such arguments for the sake of the simplicity of our article. The upper bound assumption on the density is quite natural, as this merely rules out certain degenerate distributions (i.e. when the density of some region diverges to infinity). The assumption of sub-gaussianity on the error terms is natural as we aim to analyze the least square estimate of f based on neural networks. It is well-established in the literature that for heavy-tailed errors, a least square estimator of a nonparametric function is not minimax optimal (e.g., see Han and Wellner (2019)) and one should use a variant of Huber loss function. We believe most of our analysis can be extended to this heavy-tailed error, but this is out of the scope of the current paper. The recent article (Fan et al., 2022) sheds some light on the estimation error of deep neural networks with heavy-tailed error in fixed dimension setup.

In our model (2.2), the component functions of f_0 are not identifiable without further assumptions for two reasons:

1. All the functions are identifiable up to shift, i.e. we cannot identify the difference between $(f_{0,i_1i_2}, f_{0,j_1j_2})$ and $(f_{0,i_1i_2}+c, f_{0,j_1j_2}-c)$ for a constant c.
2. The univariate marginals of the bivariate functions are not identifiable. For example, consider the subcollection of functions $\mathcal{C}_1 = \{f_{0,1}, f_{0,12}, f_{0,13}, \ldots, f_{0,1d}\}$. Let $g_k(x) = \int f_{0,1k}(x,y)\, dy$ be the marginal of $f_{0,1k}$. Then we cannot differentiate between $\mathcal{C}_1$ and a modified collection $\mathcal{C}_2 = \{f_{0,1} + \sum_{k\geq 2} g_k, f_{0,12} - g_2, \ldots, f_{0,1d} - g_d\}$.

Therefore, to identify and estimate the non-parametric functions, we need to impose certain structural conditions which will prevent us from shifting. Toward that end, we assume the following:

ASSUMPTION 2.2 (Identifiability and boundedness). *We assume the following conditions for identifiability purpose and boundedness:*

1. *All the univariate functions in* (2.2) *integrates to* 0*, i.e.* $\int_0^1 f_{0,j}(x)\, dx = 0$ *for* $1 \leq j \leq d$.
2. *All the bivariate functions have* 0 *marginals, i.e.*
$$\int_0^1 f_{0,ij}(x,y)\, dx = \int_0^1 f_{0,ij}(x,y)\, dy = 0\,.$$
3. $\|f_0\|_\infty \leq B$ *for some* $B > 0$*, where* f_0 *is defined in* (2.2).

Note that, the above assumption implies that the total integral of $f_{0,ij}$ is also 0. Assumption 2.2 ensures identifiability of all the functions involved in (2.2). Our next assumption is a smoothness assumption on the underlying component function:

ASSUMPTION 2.3. *The functions* $\{f_{0,j}\}$ *and* $\{f_{0,ij}\}$ *are assumed to be* (β, L)*-Holder smooth, i.e. the functions are* $\lfloor\beta\rfloor$ *times differentiable and* $\lfloor\beta\rfloor^{th}$ *derivative are Holder with exponent* $\beta - \lfloor\beta\rfloor$ *and constant* L.

The smoothness assumption is standard in the nonparametric regression literature, cf. (Tsybakov, 2004, Chapter 1), as this smoothness assumption controls the complexity of the underlying function class. Another common assumption is that all $f_{0,i}$ and $f_{0,ij}$ belong to a Reproducing Kernel Hilbert Space(RKHS), e.g, see Raskutti et al. (2012), Koltchinskii and Yuan (2010) and references therein. It would be interesting to study the analog of our results under such assumptions.

REMARK 2.4. *Assuming that* $\{f_{0,ij}\}$ *have a different level of smoothness* β_{ij} *is equivalent to assuming they all have* $\min_{i,j} \beta_{i,j}$ *smoothness as long as one is concerned about estimation error bounds.*

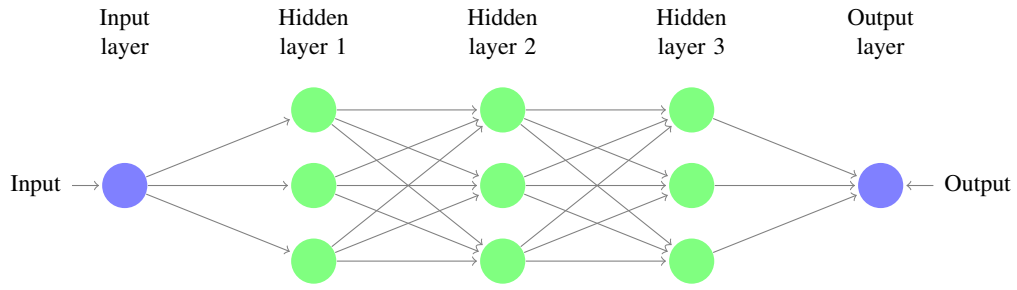

FIG 1. *A graphic representation of a deep neural network with 3 hidden layers, one input layer, one output layer.*

To estimate the mean function of (2.2) using deep neural networks, we need some properties of DNNs. A neural network is a parametric function f_θ which maps the input space to output space, i.e. in our context f_θ maps $\mathbb{R}^d$ to $\mathbb{R}$. A two-layer (i.e. one hidden layer) neural network (often termed as *shallow neural network*) with N neurons and activation function σ is defined as:

$$f_\theta(x) = a + \sum_{i=1}^{N} b_i \sigma(c_i^\top x + d_i)\,.$$

where $\theta = \{a, b_i, c_i, d_i\}$. Therefore, f_θ first projects x into $\mathbb{R}^N$ via a linear transformation $x \mapsto Cx + d$ where $c_i^\top$'s are the rows of C and d_i's are elements of d. Then it applies a non-linear activation σ to all the coordinates of $Cx + d$. Finally, it takes a linear combination of the coordinates using the map $x \mapsto a + \langle b, x\rangle$. This shallow neural network can be extended to the deep neural network by adding more hidden layers. A graphical representation of the flow of neural networks is depicted in Figure 1.

For a neural network, we denote by L the number of hidden layers (termed as depth) and by N, the maximum number of neurons at the hidden layers (denoted by width). Henceforth, we denote by $\mathcal{F}_{NN}(d, N, L, W, o)$ by the collection of all neural networks with depth L, width N, the total number of active (non-zero) weights W, input dimension d and output dimension o. We may sometimes omit the input and output dimensions in the specification of the class of neural networks when it is clear from the context.

The expressibility of neural networks has been a topic of interest for decades, which started to flourish at the beginning of 1990s. It was proved in Hornik et al. (1989) that the set of all shallow neural networks (i.e. with one hidden layer) is dense in the space of all Borel measurable functions and any Borel measurable function f can be estimated within arbitrary accuracy by increasing the number of neurons in the hidden layer. The degree of accuracy was quantified by Barron (1993) – if the Fourier transform of the function has a finite first moment, then the approximation error decreases at a rate $1/\sqrt{N}$, where N is the number of hidden neurons. The error of approximating any *smooth* function by a multi-layer feed-forward network was established by Mhaskar (1996) along with a choice of N, L which achieves the optimal accuracy. Since then, there has been a series of research on the approximation capability of deep neural networks in terms of their width, height, weights, and activation function – recently, researchers have obtained precisely bound on the approximation error of smooth functions in terms of depth and width of neural network cf. Yarotsky (2017); Lu et al. (2021); Fan and Gu (2022) and references therein. For example, Theorem 1.1 of Lu et al. (2021) shows that if $f : [0,1]^d \to \mathbb{R}$ is β times differentiable with bounded derivatives then there exists a neural network ϕ with width $c_1 N \log N$ and depth $c_2 L \log L + d$ (c_1, c_2 are some constants depending on β) with the following approximation error:

$$\|f - \phi\|_\infty \le C(NL)^{-\frac{2\beta}{d}}\,,$$

where the constant C depends on (d, β).

A related line of research delves into understanding how DNNs can successfully adapt to the unknown underlying low dimensional structure of the mean function in nonparametric regression. Performing nonparametric regression using suitable DNNs and achieving minimax optimal error bound (up to some polylog factor) initially started from the pioneering work of Kohler and Krzyżak (2005). The minimax rate of estimation of a mean function f in a standard nonparametric regression problem (e.g., additive noise model $Y = f(X) + \epsilon$) is $n^{-2\beta/(2\beta+d)}$ where β is the smoothness index of f (Assumption 2.3) and d is the dimension of X. The estimation error suffers from the curse of dimensionality, i.e. the rate is very slow for large d. However, as mentioned in the introduction, it is possible to circumvent this curse by imposing more structure on f such as additive (see Stone (1985)), or some higher order interaction model (see Stone (1994)). In Kohler and Krzyżak (2005), the authors show that it is possible to adapt to the rate $n^{-2\beta/(2\beta+k)}$ using neural network-based estimate if the model is a k-way interaction model. The key advantage of using a neural network is that we only need to specify its architecture, i.e. the width and layer, not the exact structure of the mean function and the estimated NN becomes minimax rate optimal. Recently, in a series of papers (Kohler and Krzyżak, 2016; Schmidt-Hieber, 2020; Bauer and Kohler, 2019; Kohler and Langer, 2021; Fan et al., 2022; Fan and Gu, 2022), the authors also establish that it is possible to estimate the underlying mean function f_0 at a rate $n^{-2\beta/(2\beta+k)}$ via neural networks when f_0 belongs to a more general class than k-way interaction model, called *generalized hierarchical interaction model*.

All the above analyses assume that the underlying dimension of d is fixed. This is at odds with many modern applications. Often time the constant in front of the rate of estimation depends on the dimension of X even for the simple additive model as we need to estimate d univariate functions. When d is fixed, it is possible to get away with a cruder bound. However, if $d \to \infty$, we need new techniques to obtain the optimal dependence on d, which is one of the primary goals of this paper.

We now briefly describe our estimation procedure. Consider the model in equation (2.2). We select two neural networks $(\hat{\phi}_1, \hat{\phi}_2)$ via minimizing the squared-error loss:

$$\left(\hat{\phi}_1, \hat{\phi}_2\right) = \arg\min_{\phi_1 \in \mathcal{F}^1_{NN}, \phi_2 \in \mathcal{F}^2_{NN}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \phi_1(X_i) - \phi_2(X_i))^2 \tag{2.3}$$

and set estimate $\hat{f}_{\text{big}} := \hat{\phi}_1 + \hat{\phi}_2$. Finally set the estimator $\hat{f} := sgn(\hat{f}_{\text{big}})(|\hat{f}_{\text{big}}| \wedge B)$, i.e. truncate the output of $\hat{f}$ at $[-B, B]$ as we assume $\|f_0\|_\infty \leq B$ (see point 3 of Assumption 2.2). Here the class of neural networks $\mathcal{F}^1_{NN}$ and $\mathcal{F}^2_{NN}$ are defined as:

$$\mathcal{F}^1_{NN} = \mathcal{F}_{NN}\left(d, c_1 d N_1 \log N_1, c_2 L_1 \log L_1, c_3 d (N_1 \log N_1)^2 L_1 \log L_1, 1\right) \tag{2.4}$$

$$\mathcal{F}^2_{NN} = \mathcal{F}_{NN}\left(d, c_1 d^2 N_2 \log N_2, c_2 L_2 \log L_2, c_3 d^2 (N_2 \log N_2)^2 L_2 \log L_2, 1\right) \tag{2.5}$$

for some constants c_1, c_2, c_3. For example, in $\mathcal{F}^2_{NN}$, all pairwise interactions are used as input variables. Note that we are not using a fully connected neural network here. Rather, we estimate each univariate (resp. bivariate) component function via fully-connected neural networks of width $N_1 \log N_1$ (resp. $N_2 \log N_2$) and depth $L_1 \log L_1$ (resp. $L_2 \log L_2$) and then add them. See Figure 2 for an illustration. With slight abuse of notation, we still use $\mathcal{F}^1_{NN}$ and $\mathcal{F}^2_{NN}$ to denote these specially structured neural networks. The quantities N_i and L_i typically depend on the sample size n and will be specified later. Like any other nonparametric estimator, the *generalization error* of $\hat{f}$ can be decomposed into two parts:

$$\|\hat{f} - f_0\|_{L_2(P_X)} \leq \underbrace{\|\hat{f} - f^\star\|_{L_2(P_X)}}_{\text{Statistical error}} + \underbrace{\|f^\star - f_0\|_{L_2(P_X)}}_{\text{Approximation error}} \tag{2.6}$$

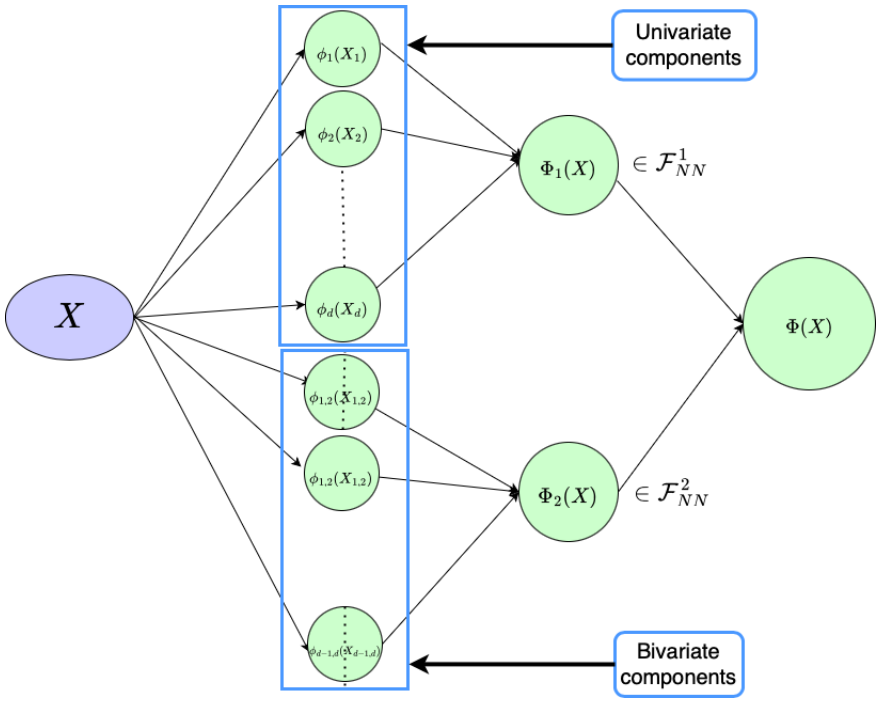

FIG 2. *A schematic diagram for the structured deep neural networks that are used to estimate the structured nonparametric regression with interactions. The first part consists of additive fully connected neural networks for the univariate predictors and the second component comprises the summation of fully connected bivariate neural networks for approximating bivariate interaction effects.*

Here $f^\star = \phi_1^\star + \phi_2^\star$ is (approximately) the best approximator of f_0 in terms of $L_2(P_X)$ norm among the class of neural networks, i.e.

$$\|f^\star - f_0\|_{L_2(P_X)} \leq \inf_{\phi_1 \in \mathcal{F}^1_{NN}, \phi_2 \in \mathcal{F}^2_{NN}} \|f_0 - \phi_1 - \phi_2\|_{L_2(P_X)} + \tau_n$$

for some small tolerance τ_n to be specified later. The rest of our analysis is divided into two parts: Subsection 2.1 bounds the approximation error and Subsection 2.2 bounds the statistical error in terms of (d, N_i, L_i, n). Finally, in Subsection 2.3 we choose the architectures of $\mathcal{F}^1_{NN}, \mathcal{F}^2_{NN}$ to balance the errors. This will prove that the rate obtained by the neural network is minimax optimal both in terms of (n, d), up to some logarithmic factor.

REMARK 2.5. *It is important to note, we are estimating f_0 as the sum of two neural networks with different architectures instead of using a single neural network. To see why this is necessary, note that the rate of convergence of the estimator is $dn^{-2\beta_1/(2\beta_1+1)}$ and $d^2 n^{-\beta_2/(\beta_2+1)}$ (Corollary 2.10) upto a logarithmic factor, where β_1 is the smoothness index of the univariate component and β_2 is the smoothness index of the bivariate components of f_0. Now note that if β_2 is sufficiently large, then $dn^{-2\beta_1/(2\beta_1+1)}$ will be dominant rate, otherwise $d^2 n^{-\beta_2/(\beta_2+1)}$ will be the dominant one. If we only use one neural network, then the rate obtained is $d^2(n^{-2\beta_1/(2\beta_1+1)} + n^{-\beta_2/(\beta_2+1)})$, which is not minimax optimal (Theorem 2.12) especially when $\beta_2 \gg \beta_1$. The trade-off between the error terms is more complicated as $d := d(n) \to \infty$ and using two neural networks (one for univariate components and the other for bivariate components) we can obtain the optimal error bounds.We conjecture that it is not possible to obtain the minimax error bound by using a single neural network.*

2.1. *Approximation error.* In this section, we compute the approximation error of mean estimation in 2-way interaction model (2.2). To this end, we invoke Theorem 1.1 of Lu et al. (2021), which states any β smooth function $f : [0,1]^d \to \mathbb{R}$ can approximate f within error of $(NL)^{-2\beta/d}$ in L_∞ norm using a DNN with width $O(N \log N)$ and depth $O(L \log L + d)$. Recall that, by Assumption, 2.2, $\int f_j(x)dx = 0$ and both the marginal integrals of f_{ij}s are 0. As stated in the previous section, we will approximate the f_is and f_{ij}s by two separate neural networks with different depths and widths. Before stating our main theorem, we first quantify a bound on the growth of d in comparison to n in the following assumption:

ASSUMPTION 2.6. *We assume that d is growing with n and*

$$\left(dn^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}} \right) \log^{4.5} n = o(1) .$$

The following theorem establishes the approximation error of the mean function via neural networks:

THEOREM 2.7. *Consider the two-way interaction model in the equation (2.2), where all the univariate components are β_1 smooth and all the bivariate components are β_2 smooth. Choose $N_1 L_1 = \lfloor n^{1/2(2\beta_1+1)} \rfloor$, $N_2 L_2 = \lfloor n^{1/2(\beta_2+1)} \rfloor$ Then we have:*

$$\inf_{\phi_2 \in \mathcal{F}^1_{NN}, \phi_2 \in \mathcal{F}^2_{NN}} \mathbb{E}\left[\|f_0(X) - \phi_1(X) - \phi_2(X)\|^2 \right] \leq C_3 \left(d(N_1 L_1)^{-4\beta_1} + \binom{d}{2} (N_2 L_2)^{-2\beta_2} \right) ,$$

where $\mathcal{F}^1_{NN}$ and $\mathcal{F}^2_{NN}$ are same as defined in (2.4) and (2.5). All the constants (C_1, C_2, C_3) are independent of (d, N_1, L_1, N_2, L_2).

The proof of Theorem 2.7 is deferred to Section 5. Here we provide the sketch of the proof. The key idea in our proof is to construct a neural network that not only approximates f but also (approximately) satisfies the identifiability constraint Assumption 2.2, i.e. the marginals of f_{ij}'s are 0. This will nullify the effect of the higher-order terms yielding the optimal rate of approximation. For univariate components, this integral constraint can be satisfied exactly; consider a function $f : [0,1] \to \mathbb{R}$ where f satisfies Assumption 2.3 with some β and $\int_0^1 f(x)\, dx = 0$. Then by (Lu et al., 2021, Theorem 1.1), there exists a neural network ϕ with width of $O(N_1 \log N_1)$ and depth $O(L_1 \log L_1)$ such that

$$\|f - \phi\|_\infty \leq C(N_1 L_1)^{-2\beta}$$

where the constant C depends only on β and Sobolev norm of f. Define $I_\phi := \int_0^1 \phi(x)\, dx$ and define a new neural network $\tilde{\phi} = \phi - I_\phi$ guaranteeing $\int \tilde{\phi} = 0$. As subtracting a constant from a neural network amounts to changing the bias of the last layer, the architecture of ϕ and $\tilde{\phi}$ are the same. Using triangle inequality,

$$\|f - \tilde{\phi}\|_\infty \leq 2C(N_1 L_1)^{-2\beta} .$$

The above idea doesn't work for the approximation of the bivariate functions f_{ij} as we not only need the integral to be 0 but also the marginals to be 0, i.e. $\int f_{ij}(x,y)\, dx = \int f_{ij}(x,y)\, dy = 0$. To address this issue, we devise a debiasing technique summarized as follows: given any bivariate function f which is β times differentiable with bounded derivatives, there exists a neural network $\phi(x,y)$ (using (Lu et al., 2021, Theorem 1.1)) with architecture $O(N_2 \log N_2), O(L_2 \log L_2)$ that satisfies:

$$\|f - \phi\|_\infty \leq C(N_2 L_2)^{-\beta} .$$

Given such a neural network ϕ, we construct ϕ_1 (resp. ϕ_2) such that $\phi_1(x)$ (resp. $\phi_2(y)$) is a neural network with width $O(N_2 \log N_2)$ and depth $O(L_2 \log L_2)$ which approximates $\int \phi(x,y)\, dy$ (resp. $\int \phi(x,y)\, dx$). Our final estimator is of the form $\tilde{\phi} = \phi - \phi_1 - \phi_2 + \iint \phi(x,y)\, dx\, dy$. We show that $\|\tilde{\phi} - f\|_\infty = O((NL)^{-\beta})$ and $\int \tilde{\phi}(x)\, dx = \int \tilde{\phi}(x,y)\, dy = O\left((NL)^{-2\beta}\right)$. This smaller approximation error ($(N_2L_2)^{-2\beta}$ instead of $(N_2L_2)^{-\beta}$) is crucial to obtain the optimal error bounds. Details of the proof can be found in Appendix 5.

2.2. *Statistical error and main theorem.* In this section, we discuss the techniques to bound the statistical error of estimation and also present the main theorem that bounds the out-of-sample prediction error. Statistical error of an estimator typically relies on the number of samples and the complexity of the underlying class of function. Recall from (2.3) that we approximate f_0 by the sum of two neural networks $\hat{\phi}_1 + \hat{\phi}_2$ where $\hat{\phi}_1$ approximates the univariate component and $\hat{\phi}_2$ approximates the bivariate component. It is immediate from the proof of Theorem 2.7 that, the approximation error depends on N and L through their product NL. Therefore, we work with the neural network with constant depth and optimize over width. So, $\mathcal{F}^1_{NN}$ and $\mathcal{F}^2_{NN}$ (equation (2.4) and (2.5)) can be revised as:

$$\mathcal{F}^1_{NN} = \mathcal{F}_{NN}\left(d, c_1 d N_1 \log N_1, c_2, c_3 d (N_1 \log N_1)^2, 1\right), \tag{2.7}$$

$$\mathcal{F}^2_{NN} = \mathcal{F}_{NN}\left(d, c_1 \binom{d}{2} N_2 \log N_2, c_2, c_3 \binom{d}{2} (N_2 \log N_2)^2, 1\right). \tag{2.8}$$

In practice, one estimates $\hat{\phi}_1, \hat{\phi}_2$ by optimizing over the weights of the neural network through gradient descent (e.g. see 6.2 of Goodfellow et al. (2016)). Since our problem is highly non-convex in terms of the weights of the neural network, it is not immediately clear whether such an algorithm will converge to global minima. To avoid such issues, we work with a neural network that minimizes empirical risk in our article. Understanding the properties of the gradient descent-based estimator is left open for future research.

The aim of this section is to bound the *variance* term, i.e. $\mathbb{E}[(\hat{f}(X) - \phi^\star(X))^2 \mid \mathbf{S}_n]$ where $\phi^\star$ is the best approximator of the mean function f_0 and $\mathbf{S}_n := \{(X_i, Y_i), i \in [n]\}$ be the training sample. The prediction/test error depends on the trade-off between the complexity of the underlying function class and the number of training samples. If the underlying function class is too complex compared to the number of samples n, then it may overfit the data resulting in high generalization error. Therefore, it is imperative to quantify the complexity of the class of neural networks. The following lemma of Bartlett et al. (2019) establishes a bound on the VC dimension of the class of neural networks with the number of active weights W and depth L:

LEMMA 2.8 ((5), Bartlett et al. (2019)). *Suppose $V(W, L)$ denotes the largest VC- dimension of a ReLU network with W parameters and L layers. There exist constants $c_1, c_2 > 0$ such that*

$$c_1 W L \log(WL) \le V(W, L) \le c_2 W L \log WL.$$

Roughly speaking, if we have a fully connected neural network with width N and depth L, then $W \sim N^2 L$ and consequently the VC dimension is $\sim N^2 L^2$. Once we have a bound on the complexity of the underlying function class, we can use techniques from empirical process theory to obtain the rate of convergence of the generalization error. For a VC-type function class with VC dimension V, the statistical error is $O(\sqrt{V/n})$ up to logarithmic factors. To see the necessity of the assumption, consider the rate obtained in Corollary 2.10. Assumption 2.6 ensures the consistency of the estimator. On the other hand, it is also necessary up to the logarithmic factor as we establish in Section 4 that the rate obtained in Corollary 2.10

is minimax optimal up to the logarithmic factor (Theorem 2.12). Therefore the assumption is the weakest possible assumption on the growth of d with respect to n under no further structural assumption. The following theorem presents a bound on the overall approximation error, combining both statistical and approximation errors.

THEOREM 2.9 (Main theorem). *Under Assumption 2.1 - 2.3 and 2.6, the ERM estimator $\hat{f}$ of f_0 defined in equation (2.3) satisfies:*

$$\mathbb{E}\left[\left(\hat{f}(X) - f_0(X)\right)^2 \mid \mathbf{S}_n\right] = O_p\left(\rho_n^2 + \frac{V_n}{n}\log^{3/2} n\right),$$

where

$$\rho_n^2 = \textit{approximation error} \leq C_1\left(dN_1^{-4\beta_1} + \binom{d}{2}N_2^{-2\beta_2}\right)$$

$$V_n = \textit{complexity} \leq C_2\left(dN_1^2\log^2 N_1 \log\left(dN_1\right) + \binom{d}{2}N_2^2\log^2 N_2\log\left(dN_2\right)\right).$$

The proof of the theorem can be found in Appendix A.1. Now we need to choose the value N_1, N_2 carefully to balance both the approximation error and statistical error. This leads to the following corollary:

COROLLARY 2.10. *Choosing $N_1 = n^{1/2(2\beta_1+1)}$ and $N_2 = n^{1/2(\beta_2+1)}$ (take the nearest integer if they are not integers) in Theorem 2.9, we have:*

$$\mathbb{E}\left[\left(\hat{f}(X) - f_0(X)\right)^2 \mid \mathbf{S}_n\right] = O_p\left(\left(dn^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2}n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^{4.5} n\right).$$

In the subsequent section, we show that this rate is minimax optimal up to the logarithmic factor. An immediate question that stems from this analysis is whether we can improve the dependence on the logarithmic factor. The recent article (Fan and Gu, 2022) proposes a slightly different approximation technique than that of Lu et al. (2021), which might reduce the power of the logarithmic factor from 4.5. However, the more interesting question is whether we can completely get rid of the log factor which is still open.

REMARK 2.11. *[Extension for k-way interaction model] Our analysis for the two-way interaction model can also be extended verbatim to the k-way interaction model. Here we provide a sketch of the extension. For identifiability of the model, we need to assume all marginals of j-component functions are 0, $j \geq 2$. We estimate f_0 via the sum of k neural networks $\phi_1 + \cdots + \phi_k$ where*

$$\phi_i \in \mathcal{F}_{NN}\left(d, c_1\binom{d}{i}N_i\log N_i, c_2, \binom{d}{i}(N_i\log N_i)^2, 1\right).$$

Since the debiasing technique used in our proofs rely on the approximation of polynomials by neural networks, they can be extended to a general k-way interaction model. This in turn controls the approximation error. The statistical error can be bounded similarly by the VC-dimension of the neural networks via Lemma 2.8.

2.3. *Minimax lower bound.* In this section, we establish the minimax lower bound for estimating the non-parametric 2-way interaction model (2.2), when the dimension of the covariate $d = d(n) \to \infty, d = o(n)$ and satisfies Assumption 2.6. For our estimation problem, the minimax risk is defined as:

$$\mathfrak{M}(n, d, \mathcal{F}) = \inf_{\hat{f}} \sup_{\substack{f \in \mathcal{F} \\ X \sim P_X}} \mathbb{E}_f \left[\left(\hat{f}(X) - f(X) \right)^2 \right]$$

Here $f(X) = \sum_{j=1}^{d} f_j(X_j) + \sum_{i<j} f_{ij}(X_i, X_j)$ and the supremum is taken over all $\{f_j\}$ and $\{f_{ij}\}$ where the component functions belongs $\Sigma(\beta, L)$ (see Assumption 2.3).

The main difficulty in establishing the minimax lower bound is to incorporate the effect of growing dimension d. Our problem is certainly harder than the problem of estimating $f_1(x_1) + f_{12}(x_1, x_2)$, assuming all other components are known in advance. The issue is how other components contribute to the lower bound in the growing d setting. The following theorem gives a precise answer.

THEOREM 2.12. *Consider the two-way interaction model as defined in equation (2.2) where each component functions $\{f_i\}_{i=1}^d$ and $\{f_{ij}\}_{1 \le i < j \le d}$ belongs to $\Sigma(\beta, L)$ (see Assumption 2.3) and denote this collection of mean functions as $\mathcal{F}$. Then the minimax rate of estimation under Assumptions 2.1-2.3 is:*

$$\mathfrak{M}(n, d, \mathcal{F}) = \inf_{\hat{f}} \sup_{\substack{f \in \mathcal{F} \\ X \sim P_X}} \mathbb{E}_f \left[\left(\hat{f}(X) - f(X) \right)^2 \right] \ge c \left(d n^{-\frac{2\beta_1}{2\beta_1 + 1}} + \binom{d}{2} n^{-\frac{2\beta_2}{2\beta_2 + 2}} \right) .$$

for some constant C independent of (n, d).

The proof of this theorem can be found in Appendix A.2. A few remarks are in order; first of all, Theorem 2.12 establishes the fact that neural network-based estimate of the mean function f_0 is minimax rate optimal up to a poly-log factor. This result complements the one in Kohler and Langer (2021); Schmidt-Hieber (2020) for the non-parametric regression in a fixed dimension regime. Secondly, Theorem 2.12 is derived under Assumption 2.3, where we assume all univariate components are β_1-smooth and all bivariate components are β_2-smooth. However, our proof can be easily adapted to the setting where each function has different smoothness and we have to pay the price for the lowest smoothness. See Remark 2.11 where we have pointed out the key steps to prove such extension.

REMARK 2.13. *[Extension for k-way interaction model] The proof of the minimax lower bound for the two-way interaction model can be easily generalized for the general $k-$way interaction model by constructing alternatives similarly and using Fano's inequality. The optimal rate is given by*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}, X \sim P_X} \mathbb{E}_{P_X} [(\hat{f}(X) - f(X))^2] \gtrsim \left(\sum_{j=0}^{k} \binom{d}{j} n^{-\frac{2\beta_j}{2\beta_j + j}} \right)$$

REMARK 2.14. *We have studied the effect of dimensionality in the minimax upper and lower bound for the interaction model in our article. A natural next step is to find Pinsker's constant for the k-way interaction model in diverging dimensions, which is left as a future research direction.*

3. Analysis of DNN when $d \gg n$. This section presents our analysis when the underlying dimension of X is larger than the sample size. Although simpler parametric models have been well-studied to avoid the curse of dimensionality, the literature on nonparametric estimation for k-way interaction model (2.1) in this regime is relatively sparse. In the nonparametric framework, a significant amount of research has been done on a high dimensional additive model under sparsity constraint where the model under consideration is as follows:

$$Y_i = \sum_{j \in S} f_{0,j}(X_{ij}) + \epsilon_i$$

where $S \subseteq \{1, \ldots, d\}$ with $s = |S| \ll n$. This is a standard assumption in high dimensional statistical analysis, where the true signal depends on a few covariates, but the active set S is apriori unknown. The analysis of the sparse additive model in high dimension was initiated in Lin and Zhang (2006) and later refined in a series of papers literature (Koltchinskii and Yuan, 2010; Ravikumar et al., 2009; Tan and Zhang, 2019; Yuan and Zhou, 2016). Koltchinskii and Yuan (2010) obtained error bounds under a global boundedness assumption, which was subsequently removed by Raskutti et al. (2012). Recently, Tan and Zhang (2019) obtained minimax guarantees when the underlying component functions lie in a reproducing kernel Hilbert space. All the previous works typically use two penalties for optimal estimation: one to control the complexity of the underlying function class and the other to control the sparsity. Yet, it is enough to use only one penalty for inducing sparsity for a neural network-based estimator, as the complexity can be controlled through the network's width and depth.

In spite of such extensive works on sparse linear models, the study of the high-dimensional linear models with k-way interaction ($k > 1$) is very sparse, letting alone k-way nonparametric interactions. For some references about the variable selection in high dimensional linear interaction models, readers may consult Zhao et al. (2009), Bien et al. (2013), Hao and Zhang (2014) and references therein. In this section, we bridge this gap by studying the asymptotic properties of the neural network estimator.

As before, we only elaborate on the analysis of the 2-way interaction model and comment on how to extend our analysis for the general k-way interaction model. The two-way sparse additive model is defined as follows:

$$Y_i = \sum_{j \in S_1} f_{0,j}(X_{ij}) + \sum_{(k,l) \in S_2} f_{0,kl}(X_{ik}, X_{il}) + \varepsilon_i \equiv f_0(X_i) + \varepsilon_i, \tag{3.1}$$

where $S_1 \subset [d]$, $S_2 \subset [d] \times [d]$ with S_1, S_2 sparse, i.e., $s_i := |S_i| \ll n$ for $i = 1, 2$. Define this class of functions by $\mathcal{F}_{\text{sp}}$. Here also we assume that the univariate components are β_1 smooth and the bivariate components are β_2 smooth. From Theorem 2.9, there exists $\{\phi_j^\star\}_{j \in S_1}, \{\phi_{kl}^\star\}_{(j,k) \in S_2}$, $j \in S_1, (j,k) \in S_2$, such that, $\phi_j^\star \in \mathcal{F}_{NN,1}$, $\phi_{jk}^\star \in \mathcal{F}_{NN,2}$ and

$$\|f_{0,j} - \phi_j^\star\|_\infty \leq C_1 N_1^{-2\beta_1} \; \forall \; j \in S_1 \,,$$

$$\|f_{0,kl} - \phi_{kl}^\star\|_\infty \leq C_2 N_2^{-\beta_2} \; \forall \; (k < l) \in S_2 \,.$$

where

$$\mathcal{F}_{NN,1} = \mathcal{F}_{NN}(1, c_1 N_1 \log N_1, c_2, c_3 N_1^2 \log^2 N_1, 1) \tag{3.2}$$

$$\mathcal{F}_{NN,2} = \mathcal{F}_{NN}(2, c_4 N_2 \log N_2, c_5, c_6 N_2^2 \log^2 N_2, 1) \,. \tag{3.3}$$

Finally, define

$$\phi^\star = \sum_{j \in S_1} \phi_j^\star + \sum_{(k<l) \in S_2} \phi_{kl}^\star. \tag{3.4}$$

Following Assumption 2.6, we assume the following growth condition on the sparsity parameter s_1, s_2:

ASSUMPTION 3.1. *The sparsity parameters s_1, s_2 corresponding to the univariate components and bivariate components satisfy the following condition:*

$$s_1 \left(n^{-\frac{2\beta_1}{2\beta_1+1}} \log^4 n + \frac{\log d}{n} \right) + s_2 \left(n^{-\frac{\beta_2}{\beta_2+1}} \log^4 n + \frac{\log d}{n} \right) = o(1) .$$

The reason behind Assumption 3.1 is similar for that of Assumption 2.6, i.e. without this assumption, there will be no consistent estimator (modulo the logarithmic factor) as it is minimax optimal rate of estimation. The analysis for the high dimensional model is not a straightforward extension of the techniques used when $d = o(n)$. To understand why, recall that even in a simple sparse linear model with ℓ_1 penalty, it is not possible to obtain a minimax rate optimal estimator without a form of *restricted strong convexity* assumption, which is typically not needed when the dimension grows slower than the sample size. Therefore, for ease of presentation, we divide the entire analysis into three parts: Section 3.1 establishes the rate of convergence for the fixed design model, i.e. when X_i's are some fixed points in $\mathbb{R}^p$. Section 3.2 deals with the random design, i.e. X is assumed to be a random variable. Finally, in Section 3.3 we present the minimax lower bound to establish that our neural network-based estimator is minimax rate optimal up to log factors.

3.1. *Analysis of Fixed design model.* In this section, we present our analysis for the fixed design model, where we assume X_i's to be fixed. Similar to (2.3), we estimate univariate and bivariate components using neural networks with different architecture, but at the same time, we use the ℓ_1-norm of $L_2(\mathbb{P}_n)$ penalty to enforce sparsity. See page 966 of Antoniadis and Fan (2001) and Yuan and Lin (2006) for such an idea in selecting a group of variables. Our estimator $\hat{f} = \sum_j \hat{\phi}_j + \sum_{k<l} \hat{\phi}_{kl}$ where components are defined as

$$\left\{\hat{\phi}_j\right\}, \left\{\hat{\phi}_{kl}\right\} = \underset{\substack{\phi_j \in \mathcal{F}_{NN,1} \\ \phi_{kl} \in \mathcal{F}_{NN,2}}}{\arg\min} \left[\frac{1}{n} \sum_i \left(Y_i - \sum_j \phi_j(X_{ij}) - \sum_{k<l} \phi_{kl}(X_{ik}, X_{il}) \right)^2 \right.$$
$$\left. + \left(\lambda_{n,1} \sum_j \|\phi_j\|_n + \lambda_{n,2} \sum_{k<l} \|\phi_{kl}\|_n \right) \right] \tag{3.5}$$

where $\mathcal{F}_{NN,1}$ and $\mathcal{F}_{NN,2}$ are defined in (3.2) and (3.3) respectively, $\|\cdot\|_n$ is the $L_2(\mathbb{P}_n)$ norm and $\lambda_{n,1}, \lambda_{n,2}$ to be specified later.

We now make couple of remarks on the estimation procedure. First, although we put $L_2(\mathbb{P}_n)$ penalty on the component functions to enforce sparsity, we may use the ℓ_1 penalty on the last layer of the weights (which adds the component neural networks). Such a penalty, in spite of its computational benefit, is difficult to analyze theoretically. Second, as we do not necessarily assume $\beta_1 = \beta_2$, we need two regularization parameters $\lambda_{n,1}$ and $\lambda_{n,2}$. We also need different architectures to adapt to the smoothness as highlighted by Remark 2.5. Furthermore, for technical simplicity, we assume the following l^∞ bound:

$$\|f_j\|_\infty \le B, \qquad \|f_{jk}\|_\infty \le B. \tag{3.6}$$

Consequently, for numerical stability, we truncate the component neural networks (i.e. ϕ_j, ϕ_{jk}) at level B. The following theorem establishes the rate of convergence.

THEOREM 3.2. *Under Assumption 2.1-2.3 and 3.1 along with (3.6), the estimator obtained in (3.5) satisfes*

$$\|\hat{f} - f_0\|_n^2 = O_p \left(s_1 \left(\rho_{n,1}^2 + \lambda_{n,1} + \frac{\lambda_{n,1}^2}{2} \right) + s_2 \left(\rho_{n,2}^2 + \lambda_{n,2} + \frac{\lambda_{n,2}^2}{2} \right) \right)$$

where $\rho_{n,1}$ (resp. $\rho_{n,2}$) is the approximation error of the univariate (resp. bivariate) components of the mean function by neural networks, bounded by (3.8), provided that

$$\lambda_{n,1} = C_3\sqrt{\frac{V_{n,1}\log n}{n} + \frac{2\log d}{n}}, \quad \lambda_{n,2} = C_4\sqrt{\frac{V_{n,2}\log n}{n} + \frac{3\log d}{n}}, \tag{3.7}$$

with $V_{n,1} = N_1^2\log^3 N_1$ and $V_{n,2} = N_2^2\log^3 N_2$.

Let us try to understand and simplify the rate obtained by Theorem 3.2. Recall from Theorem 2.7, we have:

$$\rho_{n,1} \le C_1 N_1^{-2\beta_1}, \quad \rho_{n,2} \le C_2 N_2^{-\beta_2}, \tag{3.8}$$

for some constants $C_1, C_2 > 0$. Furthermore, it is revealed in our proof (see (A.19) and (A.20)) that $\lambda_{n,1}$ and $\lambda_{n,2}$ given by (3.7) are the optimal choices, in which $V_{n,1}$ and $V_{n,2}$ are the order of VC-dimensions for $\mathcal{F}_{NN,1}$ and $\mathcal{F}_{NN,2}$ respectively. Now, typically $\lambda_{n,i} = o(1)$ (which holds as soon as $V_{n,i} = o(n/\log n)$, a condition necessary for consistency) and consequently $\lambda_{n,i}^2 = o(\lambda_{n,i})$. Hence the rate of convergence in Theorem 3.2 can be simplified as

$$\|\hat{f} - f_0\|_n^2 = O_p\left(s_1\left(\rho_{n,1}^2 + \lambda_{n,1}\right) + s_2\left(\rho_{n,2}^2 + \lambda_{n,2}\right)\right).$$

Now we choose N_1, N_2 to balance $\rho_{n,i}^2$ and $\lambda_{n,i}$, which leads to the following corollary:

COROLLARY 3.3. *Choosing $N_1 = n^{1/(2(1+4\beta_1))}$ and $N_2 = n^{1/(2(1+2\beta_2))}$ (take the nearest integer if they are not integers), we obtain:*

$$\|\hat{f} - f_0\|_n^2 = O_p\left(s_1\left(n^{-\frac{2\beta_1}{1+4\beta_1}}\log^2 n + \sqrt{\frac{\log d}{n}}\right) + s_2\left(n^{-\frac{\beta_2}{1+2\beta_2}}\log^2 n + \sqrt{\frac{\log d}{n}}\right)\right).$$

cAs we will see in Subsection 3.3, this rate is not minimax optimal. The reason is similar to that for the high dimensional sparse regression model: if we do not assume any condition on the curvature of the loss function (e.g. restricted isometry or restricted eigenvalue type conditions) then it is not possible to obtain any minimax optimal estimator which is computable in polynomial time as proved in Zhang et al. (2014). Hence, we need to assume a form of *Restricted Strong Convexity* (RSC) to obtain a minimax optimal error bound. RSC-type assumptions were popularized by Bickel et al. (2010); Candès and Tao (2007). A similar assumption is also used in the analysis of the sparse additive regression model, e.g. see (Tan and Zhang, 2019, Assumption 3). In particular, we assume the following:

ASSUMPTION 3.4. *There exist constants $\kappa_1, \kappa_2 > 0$, such that for any function $\phi = \sum_j \phi_j + \sum_{k<l}\phi_{kl}$ that satisfies:*

$$\begin{aligned}
&\lambda_{n,1}\sum_{j\in S_1^c}\|\phi_j - \phi_j^\star\|_n + \lambda_{n,2}\sum_{(k<l)\in S_2^c}\|\phi_{kl} - \phi_{kl}^\star\|_n \\
&\le 4(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + 3\lambda_{n,1}\sum_{j\in S_1}\|\phi_j - \phi_j^\star\|_n + 3\lambda_{n,2}\sum_{(k<l)\in S_2}\|\phi_{kl} - \phi_{kl}^\star\|_n \\
&\quad + s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2
\end{aligned} \tag{3.9}$$

also satisfies:

$$\kappa_1^2\sum_{j\in S_1}\|\phi_j - \phi_j^\star\|_n^2 + \kappa_2^2\sum_{(k<l)\in S_2}\|\phi_{kl} - \phi_{kl}^\star\|_n^2 \le \|\phi - \phi^\star\|_n^2 \tag{3.10}$$

where $\phi^\star$ is defined by (3.4), $\rho_{n,1}$ and $\rho_{n,2}$ are defined in Theorem 3.2 and $\lambda_{n,1}$ and $\lambda_{n,2}$ are defined in (3.7).

Before going into further details, we first compare our assumption with the standard restricted eigenvalue (RE) condition for the high dimensional linear model Bickel et al. (2009). Roughly speaking, the RE condition assumes:

$$\sum_{j\in S}(\beta_j-\beta_{0,j})^2 \lesssim \frac{1}{n}\|X(\beta-\beta_0)\|_2^2 \tag{3.11}$$

for any β that satisfies:

$$\sum_{j\in S^c}|\beta_j| \lesssim \sum_{j\in S}|\beta_j-\beta_{0,j}|\,, \tag{3.12}$$

where S is the true active set. Now consider the nonparametric additive model, where X_j affects Y through $f_j(X_j)$. Furthermore, we typically use the sum of $L_2(\mathbb{P}_n)$ norm as a penalty function in place of L_1 norm for sparsity. Ignoring the approximation error for the time being (i.e. assuming true f_j belong to the function class over which we optimize), a natural analog of condition (3.11) is the following:

$$\sum_{j\in S}\|f_j-f_{0,j}\|_n^2 \lesssim \|\hat{f}-f_0\|_n^2 \tag{3.13}$$

for all $f=\sum_j f_j$ that satisfies (similar to (3.12))

$$\sum_{j\in S^c}\|f_j\|_n \lesssim \sum_{j\in S}\|f_j-f_{0,j}\|_n\,. \tag{3.14}$$

This is precisely what was used in (Tan and Zhang, 2019, Assumption 3) for analyzing additive models in high dimensions. Assumption 3.4 is almost the same as this one with some modification due to the fact we are approximating the component functions through neural networks and consequently we may have non-zero approximation error. Ignoring the bivariate components, (3.10) is exactly the same as (3.13) with f replaced by ϕ and f_0 replaced by $\phi^\star$, which is the best approximator of f_0 from the class of neural networks. The only difference is in (3.9), which differs from (3.14) as we need to take into account the approximation error. This is precisely why the term $s_1\rho_{n,1}^2+s_2\rho_{n,2}^2$ appears on the right-hand side of equation (3.9). The other additional term $s_1\lambda_{n,1}^2+s_2\lambda_{n,2}^2$ is due to some mathematical artifact of the proof and also can be made arbitrarily smaller in order. For simplicity, if we ignore the bivariate components and this additional term for the time being, then (3.9) simplifies to:

$$\lambda_{n,1}\sum_{j\in S_1^c}\|\phi_j-\phi_j^\star\|_n \le 4s_1\rho_{n,1}^2+3\lambda_{n,1}\sum_{j\in S_1}\|\phi_j-\phi_j^\star\|_n\,. \tag{3.15}$$

If the function class is well-specified (as in the case of linear regression or additive models in Tan and Zhang (2019)), then the first term of the RHS can be removed. Consequently, cancelling $\lambda_{n,1}$ from both sides, our condition becomes same as (3.13) as used in Tan and Zhang (2019). We now state the statistical error theorem:

THEOREM 3.5. *Consider the function $\hat{\phi}$ defined by (3.5). Then, under the same assumptions as that of Theorem 3.2 along with Assumption 3.4, we have:*

$$\|\hat{f}-f_0\|_n^2 = O_p\left(s_1(\rho_{n,1}^2+\lambda_{n,1}^2)+s_2(\rho_{n,2}^2+\lambda_{n,2}^2)\right)\,. \tag{3.16}$$

The proof of this theorem is presented in Appendix A.3. The key difference between the rate obtained in Theorem 3.2 and Theorem 3.5 is the absence of $\lambda_{n,i}$, which yields a faster rate under Assumption 3.4. Similar discussion as that of after Theorem 3.2 is in order; we need to choose N_1, N_2 to balance $\rho_{n,i}^2$ and $\lambda_{n,i}^2$. This leads to the following corollary:

COROLLARY 3.6. *Choosing* $N_1 = n^{1/(2(1+2\beta_1))}$ *and* $N_2 = n^{1/(2(1+\beta_2))}$ *(take the nearest integer if they are not integers), we obtain:*

$$\|\hat{f} - f_0\|_n^2 = O_p\left(s_1\left(n^{-\frac{2\beta_1}{1+2\beta_1}}\log^4 n + \frac{\log d}{n}\right) + s_2\left(n^{-\frac{\beta_2}{1+\beta_2}}\log^4 n + \frac{\log d}{n}\right)\right).$$

In Section 3.3, we prove a matching lower bound to obtain that this rate is in fact minimax optimal up to log factors.

3.2. *Random design setting.* In the previous section, we have analyzed the fixed design model, i.e. X_i's are assumed to be fixed. Here, we extend our analysis to the random design model, namely when X_i's are random variables with distribution satisfying Assumption 2.1. As X_i's are assumed to be random, we analyze the behavior of an estimator in terms of expected squared error loss, also known as generalization/out-of-sample error. Let us first highlight the difference between this subsection and the previous one. In the fixed design setup, the entire analysis hinges on the given $(X_1, \ldots, X_n)$, i.e. we don't require to evaluate the performance of the predictor on some unobserved X. However, when we assume X's are random, then it is imperative to evaluate the performance on the unseen X's as there is a high probability (in fact probability is 1 if some component of X is continuous) that in future we need to predict on new observations. Hence, in this setup we need to bound the expected squared error loss, i.e. $\|\hat{f} - f_0\|_{L_2(P_X)}^2$ instead of empirical squared error loss, i.e. $\|\hat{f} - f_0\|_n^2$.

A key step to bound this generalization error is to control the fluctuation of the corresponding empirical process, which ensures that a predictor, that performs well on the training samples, also performs well on the previously unobserved test samples. Typically, it is difficult to achieve such a guarantee using only the L_1-norm of the $L_2(\mathbb{P}_n)$ penalties as this penalty only controls the complexity of the functions on the observed samples. Although it might be possible to obtain a minimax optimal estimator by using L_∞ penalty along with the L_1-norm of the $L_2(\mathbb{P}_n)$ penalties on the component functions, the optimization procedure becomes computationally challenging. To circumvent this issue, we propose a computationally efficient two-step procedure, which, at the cost of mildly stronger assumptions, allows us to estimate the true mean function optimally. We illustrate our estimator via Algorithm 1.

The key idea is as follows: we first split the samples into two (almost) equal halves. Based on the first half of the sample, we estimate the component functions by the same penalized procedure used in the fixed design setting (see(3.5)). Next, we use hard thresholding on the estimators of univariate and bivariate component functions to estimate the active set (Step 3 of Algorithm 1). Denote the estimated active sets by $\hat{S}_1$ and $\hat{S}_2$ respectively. Note that our threshold levels are proportional to the penalty applied in (3.5). The constants c_1, c_2 will be mentioned explicitly in the proof. Once we have $\hat{S}_1, \hat{S}_2$, then we solve the unpenalized least square problem *only on the active set* based on the second half of the data (Step 4 of Algorithm 1) to obtain the final estimate.

Our algorithm is primarily motivated by SURE independent screening (Fan and Lv, 2008; Fan and Song, 2010) and the idea of least square regression upon selecting an active set via LASSO or other penalized regression (Belloni and Chernozhukov, 2013). SURE independent screening selects the active variables by performing marginal regression for additive models, whereas for sparse linear regression, Belloni and Chernozhukov (2013) proposes to choose the active set using LASSO and then perform OLS on the selected subset to reduce bias. We combine these two ideas in Algorithm 1. First, we estimate the active subset of univariate and bivariate components $\hat{S}_1, \hat{S}_2$ (step 2-3 of Algorithm 1) and ensure that with high probability: i) $S_i \subseteq \hat{S}_i$ and ii) $|\hat{S}_i \cap S_i^c| = O(|S_i|)$.

We need i) to avoid any potential bias that may occur by ignoring active components and ii) to guard against false positives, i.e. we do not select too many inactive variables. To ensure this, we need the following assumption:

Algorithm 1: Estimation under random design setting

Input: Constant c_1, c_2 and penalty parameters $\lambda_{n,1}$, $\lambda_{n,2}$.

Output: $\hat{f}$: the estimator based on neural network.

Data: Dataset $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

1 Divide the dataset into two equal halves with $n/2$ data in each set (if n is odd, then take $(n+1)/2$ data in the first half and $(n-1)/2$ in the second half). Denote the halves by $\mathcal{D}_1$ and $\mathcal{D}_2$.

2 Using $\mathcal{D}_1$, estimate the component functions $\{\hat{\phi}_j^{\text{init}}\}$ and $\{\hat{\phi}_{jk}^{\text{init}}\}$ by solving (3.5).

3 Set $\hat{S}_1 = \{j : \|\hat{\phi}_j^{\text{init}}\|_n \geq c_1 \lambda_{n,1}\}$ and $\hat{S}_2 = \{(j<k) : \|\hat{\phi}_{jk}^{\text{init}}\|_n \geq c_2 \lambda_{n,2}\}$.

4 Re-estimate component functions on the estimated active set $\hat{S}_1$ and $\hat{S}_2$ by minimizing (un-penalized) squared error loss:

$$\{\hat{\phi}_j^{\text{final}}\}_{j\in\hat{S}_1}, \{\hat{\phi}_{jk}^{\text{final}}\}_{(j,k)\in\hat{S}_2} = \arg\min_{\substack{\phi_j\in\mathcal{F}_{NN,1} \\ \phi_{j,k}\in\mathcal{F}_{NN,2}}} \frac{1}{n}\sum_i \left(Y_i - \sum_{j\in\hat{S}_1}\phi_j(X_{ij}) - \sum_{(j,k)\in\hat{S}_2}\phi_{jk}(X_{ij}, X_{ik})\right)^2 . \tag{3.17}$$

5 Return the final estimate $\hat{f}^{\text{final}} = \sum_{j\in\hat{S}_1}\hat{\phi}_j^{\text{final}} + \sum_{(j<k)\in\hat{S}_2}\hat{\phi}_{jk}^{\text{final}}$.

ASSUMPTION 3.7. *Assume that all the component functions has minimal signal strength r_n, i.e. $\min_{j\in\mathcal{S}_1}\|f_j^0\|_2 \geq r_n$, $\min_{(j,k)\in\mathcal{S}_2}\|f_{j,k}^0\|_2 \geq r_n$ where:*

$$r_n \gtrsim \sqrt{\left(s_1\left(n^{-\frac{2\beta_1}{1+2\beta_1}}\log^4 n + \frac{\log d}{n}\right) + s_2\left(n^{-\frac{\beta_2}{1+\beta_2}}\log^4 n + \frac{\log d}{n}\right)\right)} .$$

Note that r_n is precisely the rate of convergence obtained in Corollary 3.6. Assumption 3.7 ensures that the signal in the active components is strong enough to be detected with high probability (ensuring the true active set is a subset of the estimated active set). Furthermore, by choosing an optimal penalty level $\lambda_{n,i}$ we control the number of false positives. Finally, in step 4 of Algorithm 1, we optimized squared error loss on the selected subset to produce the final estimate. The following theorem establishes rate of convergence of the estimator $\hat{f}_{\text{final}}$ obtained via Algorithm 1:

THEOREM 3.8. *Assume that the restricted strong convexity assumption (Assumption 3.4) holds with high probability on $\{X_1, \ldots, X_n\}$. Then, under Assumption 3.7, the estimator $\hat{f}^{\text{final}}$ satisfies, upto log-factors,*

$$\|\hat{f}^{\text{final}} - f_0\|_2^2 = O_p\left(s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2\right) .$$

where $\lambda_{n,1}$ and $\lambda_{n,2}$ are same as in Theorem 3.5.

The proof of the Theorem will be deferred to Appendix A.4.

3.3. *Minimax lower bound.* In this section, we establish the minimax lower bound on the rate of convergence of the high dimensional two-way interaction model. As mentioned previously, our proofs can be extended to a general k-way interaction model. Our main theorem is as follows:

THEOREM 3.9. *Suppose that we have n observations from model (3.1). Then under Assumptions 2.2, 2.1 we have:*

$$\inf_{\hat{f}} \sup_{\substack{f\in\mathcal{F}_{sp}\\ X\sim P_X}} \|\hat{f}-f\|_2^2 \gtrsim \left(s_1 \left(n^{-\frac{2\beta_1}{2\beta_1+1}} \vee \frac{\log(d/s_1)}{n} \right) + s_2 \left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee \frac{\log\left(d^2/s_2\right)}{n} \right) \right)$$

It is immediate from Theorem 3.9 and Theorem 3.8 that our estimator of the mean function based on neural networks is indeed minimax optimal up to log factors.

Although the proof techniques of Theorem 3.9 and Theorem 2.12 are similar, there are some important changes in the construction for the alternatives since we need to enforce sparsity here. We would also like to point out that our proof is different from that of Raskutti et al. (2012) since we do not assume that the components of the mean function belong to RKHS. Here we sketch the main idea of the proof briefly: each univariate component function is assumed to be β_1-smooth and each bivariate component function is assumed to be β_2-smooth. We first construct a collection of alternatives for these components and then we take a sparse combination of them. Finally, constructing alternatives along with a proper choice of the relevant hyper-parameters yield the lower bound. The proof is deferred to the Appendix.

4. Conclusion. Deep neural networks have achieved tremendous success in nonparametric function estimation. Yet due to the intrinsic difficulty of "curse-of-dimensionality", they can only handle nonparametric functions of low-dimension, without excessive restrictions on the function classes. This calls for low-dimensional nonparametric interaction models. At the same time, modern big data applications often involve nonparametric regression with a large number of predictors. Yet, most statistical theories on neural networks focus only on finite-dimensional regression. These give rise to the imminent need for the study of nonparametric interaction models in diverging dimensions and understanding the impact of dimensionality in such structured nonparametric models.

This paper contributes critically to understanding the performance of neural networks in low-order interaction models in diverging dimensions. An important conclusion of our study is that estimated components should have low biases in order to avoid unnecessary error accumulations and this is achieved by our newly introduced debiasing techniques. For slowly diverging dimensional problems, no additional regularization is needed. With proper debiasing,direct least-squares estimation on the structured neural networks is shown to achieve a rate of convergence that matches with a newly established minimax low bound, namely, it is optimal. In a high-dimensional setting, sparsity assumption on the interaction terms is necessary. We appeal to the penalized least-squares estimation and screening techniques (for random design) and show that the resulting procedure is minimax optimal by establishing a matching lower bound. Our results provide a comprehensive view of the performance of neural networks for the structured nonparametric model in diverging dimensions, which are critical to modern data science and necessary for neural networks to succeed.

5. Proof of Theorem 2.7. Suppose $f:[0,1]^d\to\mathbb{R}$ is β-smooth, i.e. it is $\lfloor\beta\rfloor$ times differentiable with bounded derivatives. Then by Taylor series approximation around some point x_0, we have:

$$\begin{aligned} f(\mathbf{x}) &= \sum_{|\boldsymbol{\alpha}|\le\beta} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\partial x^{\boldsymbol{\alpha}}}(\mathbf{x}-\mathbf{x}_0)^{\boldsymbol{\alpha}} + \sum_{|\boldsymbol{\alpha}|=\beta} \frac{1}{\boldsymbol{\alpha}!}\frac{\partial^{\boldsymbol{\alpha}} f(\mathbf{x}_0+\lambda_{\boldsymbol{\alpha}}\mathbf{x})}{\partial x^{\boldsymbol{\alpha}}}(\mathbf{x}-\mathbf{x}_0)^{\boldsymbol{\alpha}} \\ &:= T(\mathbf{x}) + R(\mathbf{x}) \end{aligned}$$

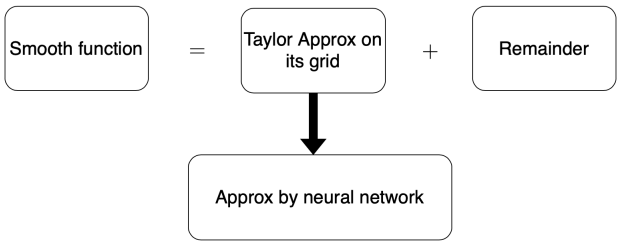

FIG 3. *Approximating smooth function by neural network*

where we use standard multi-dimensional Taylor series notation: for any $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d)$, set $|\boldsymbol{\alpha}| = \sum_{j=1}^{d} \boldsymbol{\alpha}_j$. Let $\partial^{\boldsymbol{\alpha}}/\partial x^{\boldsymbol{\alpha}}$ denotes $\partial^{\boldsymbol{\alpha}_1+\cdots+\boldsymbol{\alpha}_d}/\partial x_1^{\boldsymbol{\alpha}_1}\partial x_2^{\boldsymbol{\alpha}_2}\ldots\partial x_d^{\boldsymbol{\alpha}_d}$. For any vector $\mathbf{v}$, let $\mathbf{v}^{\boldsymbol{\alpha}}$ denote $v_1^{\boldsymbol{\alpha}_1}\ldots v_d^{\boldsymbol{\alpha}_d}$. The basic idea of approximating a smooth function by a neural network involves the following three key steps:

1. Divide the domain $[0,1]^d$ into small grids.
2. On each grid, perform Taylor series expansion around some fixed point, say the midpoint of the corner point of the grid.
3. Approximate the Taylor polynomials by neural networks.

Figure 3 presents a flowchart with these three steps. Following the notations of Lu et al. (2021), the best neural network approximator can be written as:

$$\phi(x) = \sum_{|\alpha|\leq\beta} \varphi\left(\frac{\phi_\alpha\left(\boldsymbol{\Psi}(x)\right)}{\alpha!}, \mathbf{P}_\alpha(x - \boldsymbol{\Psi}(x))\right) \tag{5.1}$$

where, $\boldsymbol{\Psi}(x)$ maps the point x to the corner point of its grid, ϕ_α approximates the value of the derivatives, $\mathbf{P}_\alpha$ approximates the polynomial $\mathbf{x}^{\boldsymbol{\alpha}}$ and finally φ (the outer function) approximates the operation xy by $\phi(x,y)$. Next we describe how to perform addition using neural networks. Given k neural networks each having width N and depth L, we can add them the following way: first send x to $2k$ neural networks $\{\phi_1, \phi_2, \ldots, \phi_k\}$ and $\{-\phi_1, \ldots, -\phi_k\}$. Then the final output is $\sum_{j=1}^{k}(\sigma(\phi_j(x)) - \sigma(-\phi_j(x)))$. As $\sigma(\cdot)$ is ReLU, we have the identity $\sigma(x) - \sigma(-x) = x$. See Figure 2 for a visual description. Therefore, the sum can be performed using a NN with width $2kN$ and depth $L+1$. If k is large, then one can perform the same operation via a neural network of width $kN \vee 2k$ and width $L+2$. In (5.1) the value of k is of the order β^d which is fixed, so the order of width and depth remains unchanged.

The neural network approximation, as presented in (5.1), has five sources of error (Table 5): i) E_1, the error of approximating f by its Taylor approximation on a grid, ii) E_2, the error of approximation of $\boldsymbol{\Psi}(\mathbf{x})$, iii) E_3, the error of approximation of ϕ_α, iv) E_4, the error of approximating $\mathbf{x}^{\boldsymbol{\alpha}}$ by $\mathbf{P}_\alpha$ and finally v) E_5, the error of approximating xy by $\phi(x,y)$. For any given N and L, suppose we divide $[0,1]$ into K equispaced intervals with length $1/K$ where $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$. If we use a neural network of width $C_1 N \log N$ and width $C_2 L \log L$ for some $C_1, C_2 > 0$, then we commit the following approximation error: Therefore for $\beta \geq 1$ and $d \geq 1$, the only term where the effect of dimension creeps in is the Taylor approximation error which is E_1. The NN $\boldsymbol{\Psi}$ is a step function, which $\mathbf{x}$ to the corner point of its grid without any error. ϕ_α is a point fitting function, which maps the corner point of the grid to its $\boldsymbol{\alpha}^{th}$ derivative of f and commits error of the order $(NL)^{-2\beta}$ which does not depend on its dimension. $\mathbf{P}_{\boldsymbol{\alpha}}$ is a neural network which approximates $\mathbf{x}^{\boldsymbol{\alpha}}$ and has error of the order $(N+1)^{-7\beta L}$ and so is the neural network φ which approximates the bivariate function $(x,y) \to xy$.

Errors	Order
E_1	$K^{-\beta} \sim (NL)^{-2\beta/d}$
E_2	0
E_3	$2(NL)^{-2\beta}$
E_4	$9\beta(N+1)^{-7\beta L}$
E_5	$216(N+1)^{-7\beta L}$

TABLE 1

Approximation error of neural network (cf. (Lu et al., 2021, Table 2))

Step 1: (Approximating one dimensional component) For each one dimensional component f_j, there exists a neural network ϕ_j (by the above mentioned construction) with width $C_1 N_1 \log N_1$ and depth $C_2 L_1 \log L_1$ such that, for some $C > 0$,

$$\|f_j - \phi_j\|_\infty \leq C(N_1 L_1)^{-2\beta_1}.$$

Define $I_j = \int_0^1 \phi_j(x)\, dx$ and another neural network $\widetilde{\phi}_j(x) = \phi_j(x) - I_j$. As subtracting a constant tantamounts to changing the bias of the last layer, it does not change the architecture. Therefore, the width and depth of $\widetilde{\phi}_j$ is same as ϕ_j. Furthermore we have:

$$\|f_j - \widetilde{\phi}_j\|_\infty = \left\| f_j - \int_0^1 f_j - \phi_j - \int_0^1 \phi_j \right\|_\infty \leq 2\|f_j - \phi_j\|_\infty \leq 2C(N_1 L_1)^{-2\beta_1}.$$

Step 2: (Approximating two dimensional component) Now consider a two dimensional function, say f_{12}. As before, following Lu et al. (2021), we can construct a neural network ϕ_{12} with width $O(N_2 \log N_2)$ and depth $O(L_2 \log L_2)$ such that

$$\|f_{12} - \phi_{12}\|_\infty \leq C(N_2 L_2)^{-\beta_2}. \tag{5.2}$$

However, by (5.1), ϕ_{12} does not approx f_{12} directly, it only approximates the Taylor expansion $T_{12}(\mathbf{x})$. Hence, there exists some $L^\star(\beta) > 0$ such that for any $L \geq L^\star(\beta)$,

$$\|f_{12}(\mathbf{x}) - \phi_{12}(\mathbf{x})\|_\infty \leq \underbrace{\|T_{12}(\mathbf{x}) - \phi_{12}(\mathbf{x})\|_\infty}_{O((N_1 L_1)^{-2\beta_2})} + \underbrace{\|R_{12}(\mathbf{x})\|_\infty}_{O((N_2 L_2)^{-\beta_2})}. \tag{5.3}$$

Note that $T_{12}(\mathbf{x})$ is a bivariate polynomial. Although the overall approximation error of the neural network is $O((N_2 L_2)^{-\beta_2})$, the approximation error of T_{12} by ϕ_{12} is much faster and can be assumed to be faster than $(N_2 L_2)^{-2\beta_2}$ if $L \geq L^\star(\beta_2)$ for some constant $L^\star(\beta_2)$ only depending on β_2. If we define $T_1(x)$ (resp. T_2(y)) as $T_1(x) = \int_0^1 T_{12}(x,y)\, dy$ (resp. $T_2(y) = \int_0^1 T_{12}(x,y)\, dx$), then both T_1 and T_2 are univariate polynomial of degree $\leq \beta_2 - 1$ and hence β_2 smooth. Consequently, we can construct NN ξ_1 and ξ_2 with width of $O(N_2 \log N_2)$ and depth of $O(L_2 \log L_2)$ such that:

$$\|T_1(x) - \xi_1(x)\|_\infty \leq C(N_2 L_2)^{-2\beta_2}, \tag{5.4}$$

$$\|T_2(y) - \xi_2(y)\|_\infty \leq C(N_2 L_2)^{-2\beta_2}. \tag{5.5}$$

We define the constant $I_{12} = \iint \phi_{12}(x,y)\, dx\, dy$. Finally, we define the estimator by $\widetilde{\phi}_{12}(x,y) := \phi(x,y) - \xi_1(x) - \xi_2(y) + I_{12}$. Let us comment about the architecture of $\widetilde{\phi}_{12}$. As mentioned before, adding I_{12} does not change the architecture. Also as ϕ_{12}, ξ_1, ξ_2 has width $O(N_2 \log N_2)$ and depth $O(L_2 \log L_2)$, $\widetilde{\phi}_{12}$ also has width $O(N_2 \log N_2)$ and depth $O(L_2 \log L_2)$ (see Figure 2). Next, we show that the L_∞ distance of $\widetilde{\phi}_{12}$ form f_{12} remains

of the order $(N_2L_2)^{-\beta_2}$. Recall that the marginals of f_{12} are 0. Using triangle inequality we have:

$$\begin{aligned}
\|f_{12}-\widetilde{\phi}_{12}\|_\infty &= \|f_{12}-\phi_{12}+\xi_1+\xi_2-I_{12}\|_\infty \\
&\le \|f_{12}-\phi_{12}\|_\infty + \left\|\int_0^1 f_{12}\,dy-\xi_1\right\|_\infty \\
&\qquad + \left\|\int_0^1 f_{12}\,dx-\xi_2\right\|_\infty + \left\|\iint_{[0,1]^2} f_{12}\,dx\,dy-I_{12}\right\|_\infty \\
&\le 2\|f_{12}-\phi_{12}\|_\infty + \left\|T_1+\int_0^1 R_{12}\,dy-\xi_1\right\|_\infty \\
&\qquad + \left\|T_2+\int_0^1 R_{12}\,dx-\xi_2\right\|_\infty \quad \left[\text{As } I_{12}=\iint \phi_{12}\right] \\
&\le 2\|f_{12}-\phi_{12}\|_\infty + \|T_1-\xi_1\|_\infty + \|T_2-\xi_2\|_\infty \\
&\qquad + \left\|\int_0^1 R_{12}\,dy\right\|_\infty + \left\|\int_0^1 R_{12}\,dx\right\|_\infty \\
&\le 2\|f_{12}-\phi_{12}\|_\infty + \|T_1-\xi_1\|_\infty + \|T_2-\xi_2\|_\infty + 2\|R_{12}\|_\infty \\
&\le C(N_2L_2)^{-\beta_2} + 2C(N_2L_2)^{-2\beta_2} + 2C(N_2L_2)^{-\beta_2} = 5C(N_2L_2)^{-\beta_2}\,.
\end{aligned}$$

Here the first error follows from equation (5.2), second and third from equation (5.4) and the last one from equation (5.3).

We next show that $\left\|\int_0^1\left(\widetilde{\phi}_{12}-f_{12}\right)\,dy\right\|_\infty \lesssim (NL)^{-2\beta}$. Since $\int_0^1 f_{12}\,dy=0$, the claim is true since

$$\begin{aligned}
\left\|\int_0^1 \widetilde{\phi}_{12}\,dy\right\|_\infty &= \left\|\int_0^1 \phi_{12}\,dy-\xi_1-\int_0^1\xi_2\,dy+I_{12}\right\|_\infty \\
&\le \left\|\int_0^1(\phi_{12}-T_{12})\,dy+T_1-\xi_1\right\|_\infty \\
&\qquad + \left\|\int_0^1(\xi_2-T_2)\,dy+\int_0^1 T_2\,dy-I_{12}\right\|_\infty \\
&\le \|\phi_{12}-T_{12}\|_\infty + \|\xi_1-T_1\|_\infty + \|\xi_2-T_2\|_\infty \\
&\qquad + \left\|\iint_{[0,1]^2} T_{12}\,dx\,dy-\iint_{[0,1]^2}\phi\,dx\,dy\right\|_\infty \\
&\le 2\|\phi_{12}-T_{12}\|_\infty + \|\xi_1-T_1\|_\infty + \|\xi_2-T_2\|_\infty \le 4C(N_2L_2)^{-2\beta_2}
\end{aligned}$$

Here, the first inequality follows from equation (5.3) second and third inequality follows form (5.4).

Step 3: (Combining all the bounds) Here, we combine the bounds for one dimensional and two dimensional components to prove Theorem 2.7. Define the function $\widetilde{\phi}$ as:

$$\widetilde{\phi}=\sum_{j=1}^d \widetilde{\phi}_j+\sum_{k<l}\widetilde{\phi}_{kl}\,.$$

Ging back to our argument for summing multiple neural network, the width of $\sum_j \tilde{\phi}_j$ is $O(dN_1 \log N_1)$ and depth is $O(L_1 \log L_1)$, the width of $\sum_{k<l} \widetilde{\phi}_{kl}$ is $O((d(d-1)/2)N_2 \log N_2)$ and depth is $O(L_2 \log L_2)$. Furthermore, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left(f_0(X)-\widetilde{\phi}(X)\right)^2\right] \le p_{\max}\int_{[0,1]^d}\left(f_0(\mathbf{x})-\widetilde{\phi}(\mathbf{x})\right)^2 d\mathbf{x}\\
&= p_{\max}\int_{[0,1]^d}\left(\sum_j f_{0,j}(x_j)+\sum_{i<j} f_{0,ij}(x_i,x_j)-\sum_j \widetilde{\phi}_j(x_j)-\sum_{i<j}\widetilde{\phi}_{ij}(x_i,x_j)\right)^2 d\mathbf{x}\\
&\le 2p_{\max}\int_{[0,1]^d}\left(\sum_j f_{0,j}(x_j)-\sum_j \widetilde{\phi}_j(x_j)\right)^2 d\mathbf{x} + 2p_{\max}\int_{[0,1]^d}\left(\sum_{i<j} f_{0,ij}(x_i,x_j)-\sum_{i<j}\widetilde{\phi}_{ij}(x_i,x_j)\right)^2 d\mathbf{x}\\
&= 2p_{\max}\sum_{j=1}^d\int_0^1\left(f_{0,j}(x_j)-\widetilde{\phi}_j(x_j)\right)^2 dx_j + 2p_{\max}\sum_{i<j}\int_{[0,1]^2}\left(f_{0,ij}(x_i,x_j)-\widetilde{\phi}_{ij}(x_i,x_j)\right)^2 dx_i\,dx_j\\
&+ 2p_{\max}\sum_{j\neq j'}\int_0^1\left(f_{0,j}(x_j)-\widetilde{\phi}_j(x_j)\right)dx_j\int_0^1\left(f_{0,j'}(x_{j'})-\widetilde{\phi}_{j'}(x_{j'})\right)^2 dx_{j'}\\
&+ 2p_{\max}\int_{[0,1]^4}\sum_{(i,j)\neq(k,l)}\left(f_{0,ij}(x_i,x_j)-\widetilde{\phi}_{ij}(x_i,x_j)\right)\left(f_{0,kl}(x_k,x_l)-\widetilde{\phi}_{kl}(x_k,x_l)\right)dx_i dx_j dx_k dx_l\\
&\le 2p_{\max}\left[\sum_{j=1}^d\|f_{0,j}-\widetilde{\phi}_j\|_\infty^2+\sum_{i<j}\|f_{0,ij}-\widetilde{\phi}_{ij}\|_\infty^2\right]\\
&\quad + 2p_{\max}\sum_{\substack{i\neq(k,l)\\ j\neq(k,l)}}\int_{[0,1]^2}\widetilde{\phi}_{ij}\,dx_i\,dx_j\int_{[0,1]^2}\widetilde{\phi}_{kl}\,dx_k\,dx_l +\\
&\quad + 2p_{\max}\sum_{i,j,k}\int_0^1\left(\int_0^1\widetilde{\phi}_{ij}\,dx_j\right)\left(\int_0^1\widetilde{\phi}_{ik}\,dx_k\right)dx_i\\
&\le 2C^2 p_{\max}\left[d(N_1L_1)^{-4\beta_1}+\binom{d}{2}(N_2L_2)^{-2\beta}+\binom{d}{4}(N_2L_2)^{-4\beta}+\binom{d}{3}(N_2L_2)^{-4\beta}\right]\\
&\le C_3\left(d(N_1L_1)^{-4\beta_1}+\binom{d}{2}(N_2L_2)^{-2\beta_2}\right),
\end{aligned}
$$

for some $C, C_3 > 0$ since $d^2(N_2L_2)^{-2\beta_2} \to 0$ by Assumption 2.6. Here the penultimate inequality follows by combining the results of Step 1 and Step 2. This completes the proof.

Funding. This paper is supported by ONR N00014-22-1-2340 and the NSF grants DMS-2052926, DMS-2053832, DMS-2210833.

REFERENCES

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association* **96** 939–967.

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

BARTLETT, P. L., HARVEY, N., LIAW, C. and MEHRABIAN, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research* **20** 2285–2301.

BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics* **47** 2261–2285.

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector .

BICKEL, P. J., RITOV, Y., TSYBAKOV, A. B. ET AL. (2010). Hierarchical selection of variables in sparse high-dimensional regression. *IMS Collections* **6** 28.

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *The Annals of statistics* **41** 1111.

BREIMAN, L. (2001). Random forests. *Machine learning* **45** 5–32.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* 453–510.

CANDÈS, E. and TAO, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics* **42** 1564–1597.

FAN, J. and GIJBELS, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall.

FAN, J. and GU, Y. (2022). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *arXiv preprint arXiv:2210.02002* .

FAN, J., GU, Y. and ZHOU, W.-X. (2022). How do noise tails impact on deep relu networks? *arXiv preprint arXiv:2203.10418* .

FAN, J., HÄRDLE, W. and MAMMEN, E. (1998). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics* **26** 943–971.

FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical foundations of data science*. Chapman and Hall/CRC.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* **38** 3567–3604.

FREUND, Y., SCHAPIRE, R. E. ET AL. (1996). Experiments with a new boosting algorithm. In *icml*, vol. 96. Citeseer.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *Journal of the American statistical Association* **76** 817–823.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep learning*. MIT press.

HAN, Q. and WELLNER, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics* **47** 2286–2319.

HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **109** 1285–1301.

HARDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics* **21** 157–178.

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks* **2** 359–366.

HOROWITZ, J. L. and MAMMEN, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *The Annals of Statistics* **35** 2589–2619.

KOHLER, M. and KRZYŻAK, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Statistics* **17** 891–913.

KOHLER, M. and KRZYŻAK, A. (2016). Nonparametric regression based on hierarchical interaction models. *IEEE Transactions on Information Theory* **63** 1620–1630.

KOHLER, M. and LANGER, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics* **49** 2231–2249.

KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics* **38** 3660–3695.

KOSOROK, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

KPOTUFE, S. and DASGUPTA, S. (2012). A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences* **78** 1496–1515.

KRIZHEVSKY, A., HINTON, G. ET AL. (2009). Learning multiple layers of features from tiny images .

LECUN, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* .

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34** 2272–2297.

LU, J., SHEN, Z., YANG, H. and ZHANG, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis* **53** 5465–5506.

MHASKAR, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural computation* **8** 164–177.

MUKHERJEE, D., BANERJEE, M. and RITOV, Y. (2021). Optimal linear discriminators for the discrete choice model in growing dimensions. *The Annals of Statistics* **49** 3324–3357.

NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S., NELSON, M. R. ET AL. (2008). Genes mirror geography within europe. *Nature* **456** 98–101.

RASKUTTI, G., J WAINWRIGHT, M. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of machine learning research* **13**.

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 1009–1030.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics* **48** 1875–1897.

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

STONE, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13** 689–705.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22** 118–171.

TAN, Z. and ZHANG, C.-H. (2019). Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics* **47** 2567–2600.

TSYBAKOV, A. B. (2004). Introduction to nonparametric estimation, 2009. *URL https://doi. org/10.1007/b13794. Revised and extended from the* **9**.

VAART, A. V. D. and WELLNER, J. A. (1997). Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society* **160** 596–608.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*. Springer, 16–28.

YANG, Y. and DUNSON, D. B. (2016). Bayesian manifold regression. *The Annals of Statistics* **44** 876–905.

YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics* **43** 652–674.

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

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.

YUAN, M. and ZHOU, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics* **44** 2564–2593.

ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*. PMLR.

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection .

APPENDIX A: PROOFS

A.1. Proof of Theorem 2.9. We want to invoke (Vaart and Wellner, 1997, Theorem 3.2.5). For any $\phi = \phi_1 + \phi_2$ (where $\phi_1 \in \mathcal{F}^1_{NN}, \phi_2 \in \mathcal{F}^2_{NN}$), define the population risk function $R(\phi)$ (resp. empirical risk $\hat{R}_n(\phi)$) as

$$R(\phi) = \mathbb{E}[(Y - \phi(X))^2] \quad \left(\text{resp. } \hat{R}(\phi) = \frac{1}{n} \sum_i (Y_i - \phi(X_i))^2 \right)$$

Since that ϵ is independent of X with mean 0, it is immediate that:

$$R(\phi) - R(f_0) = \mathbb{E}[(f(X) - f_0(X))^2] \triangleq d^2(\phi, f_0)\,.$$

Define $\mathcal{F}_n := \{\phi = \phi_1 + \phi_2 : \phi_1 \in \mathcal{F}^1_{NN}, \phi_2 \in \mathcal{F}^2_{NN}\}$ (i.e. $\mathcal{F}_n = \mathcal{F}^1_{NN} + \mathcal{F}^2_{NN}$) and $\omega_n = d(\phi^\star, f_0)$, i.e. the approximation error of the class of neural network. It is immediate that:

$$\mathcal{F}_n \subseteq \mathcal{F}_{NN}\left(d, c_1\left(dN_1 \log N_1 + \tbinom{d}{2} N_2 \log N_2\right), c_2, W_n, 1\right)$$

with

$$W_n = c_3\left(d(N_1 \log N_1)^2 + \binom{d}{2}(N_2 \log N_2)^2\right),$$

for some constant c_1, c_2, c_3 independent of (n, d). Therefore, VC dimension of $\mathcal{F}_n$ is less than or equal to $V_n \triangleq c_4 W_n \log W_n$ for some $c_4 > 0$. We use this in the subsequent analysis. Let $\zeta_n(\delta)$ be such that:

$$\mathbb{E}\left[\sup_{\phi \in \mathcal{F}_n : d(\phi, f_0) \le \delta} |(R_n(\phi) - R(\phi)) - (R_n(f_0) - R(f_0))|\right] \lesssim \frac{\zeta_n(\delta)}{\sqrt{n}}\,. \tag{A.1}$$

This function ζ_n is called the modulus of continuity, which is used to bound the fluctuation of the empirical process around the population process in a neighborhood of the true function. $\zeta_n(\delta)$ intrinsically depends on the complexity of the underlying function class. We quantify this dependency through the following maximal inequality (e.g. Theorem 5.2 of Chernozhukov et al. (2014)), which we state here for the convenience of the readers:

LEMMA A.1 (Theorem 5.2 of Chernozhukov et al. (2014)). *Define the operator $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. Consider a collection of functions $\mathcal{F}$ with envelope function F. Assume that $F \in L_2(P)$. Then:*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)|\right] \lesssim J(r, \mathcal{F}, F)\|F\|_2 + \frac{\|M\|_2 J^2(r, \mathcal{F}, F)}{r^2 \sqrt{n}}$$

for $r = \sigma / \|F\|_2$, where

$$J(\tau, F, \mathcal{F}) = \int_0^\tau \sup_Q \sqrt{1 + \log N\left(\epsilon \|F\|_2, \mathcal{F}, L_2(Q)\right)}\, d\epsilon$$

$$M = \max_{1 \le i \le n} F(X_i), \qquad \sigma^2 = \sup_{f \in \mathcal{F}} P f^2\,.$$

We use Lemma A.1 to find the function $\zeta_n(\delta)$. For any ϕ, define a function $g \equiv g(\phi)$ on the space of (X, ϵ) as:

$$\begin{aligned} g(X, \epsilon) &= (f_0(X) - \phi(X) + \epsilon)^2 - \epsilon^2 \\ &= (\phi(X) - f_0(X))^2 + 2(\phi(X) - f(X))\epsilon. \end{aligned} \tag{A.2}$$

Define the collection of $\mathcal{G}$ as $\mathcal{G} = \{g(\phi) : \phi \in \mathcal{F}_n\}$. This definition implies:

$$\mathbb{E}\left[\sup_{\phi: d(\phi, f_0) \le \delta} |(R_n(\phi) - R(\phi)) - (R_n(f_0) - R(f_0))|\right] = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{g \in \mathcal{G}_\delta} |\mathbb{G}_n(g)|\right].$$

Here we $\mathcal{G}_\delta$ as the collection of all $g(\phi)$'s, such that $d(\phi, f_0) \le \delta$. To apply Lemma A.1, we need to i) quantify σ^2, ii) find an envelope of $\mathcal{G}_\delta$ and iii) bound the logarithm of covering number. For i) observe that:

$$\sup_{g \in \mathcal{G}: d(f, f_0) \le \delta} \mathbb{E}[g(X)^2] \le \delta^2(2 + 8\sigma_\epsilon^2) \triangleq C_{\sigma_\epsilon}\delta^2 .$$

Next, to construct an envelope, note that (A.2) implies:

$$|g(X, \epsilon)| \le 9B^2 + 6B\epsilon .$$

for all (X, ϵ). This follows from the fact that $\|\phi\|_\infty \le 2B$ and $\|f_0\|_\infty \le B$. Therefore, we can take the envelope G as $G(X, \epsilon) = 9B^2 + 6B\epsilon$. To bound the logarithm of the covering number of $\mathcal{G}_\delta$, we use Lemma 9.9 of Kosorok (2008). Note that, each $g = g_1 + g_2 + hg_3$ where $g_1 = (\phi - f_0)_+^2$, $g_2 = (\phi - f_0)_-^2$, $g_3 = 2(\phi - f_0)$ and $h(x, \epsilon) = \epsilon$. We calculate the VC dimensions of each class of functions:

- The VC dimension of $\mathcal{G}_1 = \{g_1(\phi) : \phi \in \mathcal{F}_{NN}\}$ is $\le V_n$. To see this, write $g_1(x, \epsilon) = m_1 \circ (\phi - f_0) \circ m_2(x, \epsilon)$ where $m_2(x, \epsilon) = x$ and $m_1(x) = x_+^2$. As precomposing by a fixed function ϕ, subtracting a fixed function f_0 and post-composing by a monotone function does not change the VC dimension, the claim follows. Similar argument establishes that VC dimension of $\mathcal{G}_2 = \{g_2(f) : f \in \mathcal{F}\}$ is $\le V_n$.
- VC dimension of $\mathcal{G}_3$, the collection of $g_3 h$ is $\lesssim V_n$. To see this, note that VC dimension of g_3 is $\le V_n$. As before, $g_3 = 2(\phi - f_0) \circ \psi$ as we have already mentioned that pre-composing by ψ, subtracting f_0 and multiplying by by a fixed function (here $2h$) does not change the VC dimension.

Therefore, we have established that VC-dim of $\mathcal{G}_i$ is $\lesssim V_n$. This implies, via Hausslers's bound (Van Der Vaart and Wellner, 1996, Theorem 2.6.7)):

$$\sup_Q \log N\left(\epsilon \|G_i\|_2, \mathcal{G}_i, L_2(Q)\right) \lesssim \log KV_n + 2V_n \log\left(\frac{2\sqrt{2e}}{\epsilon}\right).$$

As $\mathcal{G}_\delta \subseteq \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3$, we have:

$$\begin{aligned}
\log N\left(\epsilon \|G\|_2, \mathcal{G}_\delta, L_2(Q)\right) &\lesssim \sum_{j=1}^{3} \log N\left(\epsilon \|G_i\|_2, \mathcal{G}_i, L_2(Q)\right) \\
&\le 3 \log KV_n + 6V_n \log\left(\frac{2\sqrt{2e}}{\epsilon}\right) \\
&\le 9V_n \log\left(\frac{2\sqrt{2e}}{\epsilon}\right),
\end{aligned}$$

where the last inequality holds as soon as $\log KV_n \le V_n$ and $\epsilon \le 2\sqrt{2/e}$. Therefore, we can use the last inequality for all small ϵ and for all large enough V_n. Taking $r = \sqrt{C_{\sigma_\epsilon}}\delta / \|G\|_2$ we have:

$$J(r, \mathcal{G}_\delta, G) \le \int_0^r \sqrt{1 + 9V_n \log\left(\frac{2\sqrt{2e}}{\epsilon}\right)}\, d\epsilon$$

$$\leq 4\sqrt{V_n}\int_0^r \sqrt{\log\left(\frac{2\sqrt{2e}}{\epsilon}\right)}\,d\epsilon$$
$$= 8\sqrt{2eV_n}\int_0^{r/2\sqrt{2e}} \sqrt{\log\left(\frac{1}{\epsilon}\right)}\,d\epsilon$$
$$\leq 4r\sqrt{V_n}\sqrt{\log\left(\frac{2\sqrt{2e}}{r}\right)}.$$

Here the second inequality holds as long as $V_n \log(2\sqrt{2}/\epsilon) \geq 1/7$, i.e. $\epsilon < 2\sqrt{2}\exp(-1/(7V_n))$. This holds for all large n as long as $r < 1$. Hence we conclude via Lemma A.1:

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_\delta}|\mathbb{G}_n(g)|\right] \lesssim \delta\sqrt{V_n\log\left(\frac{1}{\delta}\right)} + \frac{V_n}{\sqrt{n}}\log\left(\frac{1}{\delta}\right)\sqrt{\mathbb{E}\left[\max_i(1+2|\epsilon_i|)^2\right]}$$
$$\lesssim \delta\sqrt{V_n\log\left(\frac{1}{\delta}\right)} + \frac{V_n\sqrt{\log n}}{\sqrt{n}}\log\left(\frac{1}{\delta}\right) \triangleq \phi_n(\delta)\,.$$

Now we establish the rate of convergence. Set $r_n = \left(\sqrt{\frac{V_n}{n}\log\frac{n}{V_n}}\sqrt{\log n} + \omega_n\right)^{-1}$, where $\omega_n = d(\phi^\star, f_0)$ is the approximation error defined before. Note that r_n satisfies the equation $\phi_n(r_n^{-1}) \lesssim r_n^{-2}\sqrt{n}$ since

$$r_n^2\phi_n\left(\frac{1}{r_n}\right) \lesssim r_n\sqrt{V_n\log r_n} + r_n^2\frac{V_n\sqrt{\log n}}{\sqrt{n}}\log r_n$$
$$\leq \sqrt{n}\left[\left(\frac{1}{2}\frac{\log\frac{n}{V_n} - \log\log\frac{n}{V_n} - \frac{1}{2}\log\log n}{\log\frac{n}{V_n}\sqrt{\log n}}\right)^{1/2} + \left(\frac{1}{2}\frac{\log\frac{n}{V_n} - \log\log\frac{n}{V_n} - \frac{1}{2}\log\log n}{\log\frac{n}{V_n}}\right)\right]$$
$$\lesssim \sqrt{n}$$

We now use shelling argument to establish the rate of convergence. Fix $t > 1$, and define $A_j := \{\mathcal{F}_{NN} : 2^{j-1}t \leq r_n d(f_0, \phi) \leq 2^j t\}$. It follows from the definition of ERM, we have

$$\mathbb{P}\left(r_n d(\hat{\phi}, f_0) \geq t\right) = \mathbb{P}\left(\sup_{\substack{\phi\in\mathcal{F}_{NN}\\ r_n d(f_0,\phi)\geq t}} \hat{R}_n(\phi^\star) - \hat{R}_n(\phi) \geq 0\right)$$
$$\leq \sum_{j=1}^\infty \mathbb{P}\left(\sup_{\phi\in A_j} \hat{R}_n(\phi^\star) - \hat{R}_n(\phi) \geq 0\right)$$
$$\leq \sum_{j=1}^\infty \mathbb{P}\left(\sup_{\phi\in A_j} (R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi)) \geq \inf_{\phi\in A_j} R(\phi) - R(\phi^\star)\right)$$
$$\leq \sum_{j=1}^\infty \frac{\mathbb{E}\left[\sup_{\phi\in A_j}|(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi))|\right]}{\inf_{\phi\in A_j} R(\phi) - R(\phi^\star)} \tag{A.3}$$

We bound the numerator and denominator separately. For the denominator note that:

$$\inf_{f\in A_j} R(\phi) - R(\phi^\star) \geq \inf_{f: d_{f,f_0}\geq 2^{j-1}tr_n^{-1}} R(\phi) - R(f_0) + R(f_0) - R(\phi^\star)$$

$$\geq 2^{2j-2}t^2 r_n^{-2} - \omega_n^2$$

For the numerator:

$$\mathbb{E}\left[\sup_{\phi\in A_j} |(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi))|\right]$$

$$\leq \mathbb{E}\left[\sup_{f:d(f,f_0)\leq 2^j t r_n^{-1}} |(R_n(\phi^\star) - R(\phi^\star)) - (R_n(\phi) - R(\phi))|\right] \lesssim \frac{\zeta_n\left(2^j t r_n^{-1}\right)}{\sqrt{n}}\,.$$

Here, the second last equation holds because $d(\phi^\star, f_0) = \omega_n \leq 2^j t r_n^{-1}$ as $r_n^{-1} \geq \omega_n$. Putting this bound in equation (A.3) we have:

$$\begin{aligned}
\mathbb{P}\left(r_n d(\hat{f}, f_0) \geq t\right) &\leq \sum_{j=1}^{\infty} \frac{\phi_n\left(2^j t r_n^{-1}\right)}{\sqrt{n}\left(2^{2j-2}t^2 r_n^{-2} - \omega_n^2\right)} \\
&\leq \sum_{j=1}^{\infty} \frac{2^j t \phi_n\left(r_n^{-1}\right)}{\sqrt{n}\left(2^{2j-2}t^2 r_n^{-2} - \omega_n^2\right)} = \sum_{j=1}^{\infty} \frac{2^j t r_n^2 \phi_n(r_n^{-1}) n^{-1/2}}{2^{2j-2}t^2 - \omega_n^2 r_n^2} \\
&\lesssim \sum_{j=1}^{\infty} \frac{2^j t}{2^{2j-2}t^2 - 1} \qquad [\text{since } r_n^2\phi_n(r_n^{-1}) \lesssim \sqrt{n},\ \ \omega_n^2 r_n^2 \leq 1] \\
&= \frac{1}{t}\sum_{j=1}^{\infty} \frac{2^j}{2^{2j-2} - \frac{1}{t^2}} \leq \frac{c}{t}
\end{aligned}$$

This proves that:

$$r_n\, d\left(\hat{\phi}, f_0\right) = O_p(1)\,.$$

i.e.

$$\|\hat{f} - f_0\|^2_{L_2(P_X)} = O_p\left(\omega_n^2 + \frac{V_n}{n}\log^{3/2} n\right)\,.$$

Now to balance ω_n and V_n, we choose $N_1 = \lfloor n^{1/2(2\beta_1+1)} \rfloor$ and $N_2 = \lfloor n^{1/2(\beta_2+1)} \rfloor$. This implies:

$$\begin{aligned}
\omega_n^2 &\lesssim d n^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}} \\
\frac{V_n}{n} &\lesssim \frac{W}{n}\log W \lesssim \left(d n^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^3 n,
\end{aligned}$$

which yields

$$\|\hat{f} - f_0\|^2_{L_2(P_X)} = O_p\left(\left(d n^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}}\right)\log^{4.5} n\right).$$

This completes the proof.

A.2. Proof of Theorem 2.12. Here, we establish the lower bound on the following two-way interaction model:

$$Y_i = \mu + \sum_{j=1}^{d} f_j(X_{ij}) + \sum_{k<l} f_{kl}(X_{ik}, X_{il}) + \epsilon_i \triangleq f_0(X_i) + \epsilon_i\,.$$

We show that:

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}, X\sim P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq c\left(dn^{-\frac{2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{-\frac{\beta_2}{\beta_2+1}}\right).$$

where $\mathcal{F}$ is the collection of all additive functions which satisfies: I) The one-dimensional components are β_1-smooth, $[0,1]$ and integrate to 0 and II) the two-dimensional components are β_2-smooth, supported on $[0,1]^2$ having marginals 0. Our proof is based on the techniques introduced in (Tsybakov, 2004, Section 2.6.1). Throughout the proof, assume $X_{ij} \overset{\text{iid}}{\sim} Unif(0,1)$, $i \leq n$, $j \leq d$ and $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, $i \leq n$. The key idea is to use Fano's inequality upon carefully choosing a subset of $\mathcal{F}$. Following Tsybakov (2004), fix the following notations:

$$m_i = \lfloor c_0 n^{\frac{1}{2\beta_i+1}} \rfloor, \quad h_i = m_i^{-1}, \quad i = 1,2. \tag{A.4}$$

We construct the alternatives in two steps, for one-dimensional components and for two-dimensional components.

Alternatives for one-dimensional components: First divide $[0,1]$ into m_1 grids $[(k-1)/m_1, k/m_1]$, $1 \leq k \leq m_1$. For each k, define $x_k = (k-0.5)/m_1$. Select a smooth kernel $K \in C^\infty(\mathbb{R})$ supported on $(-1/2, 1/2)$, $\int K(u)\, du = 0$ and define $\int K^2(u)\, du \triangleq \|K\|_2^2$. Define the functions $\phi_k(x)$, $k \leq m_1$ as:

$$\phi_k(x) := Lh_1^{\beta_1} K\left(\frac{x - x_k}{h_1}\right) \tag{A.5}$$

The constant L is chosen so that ϕ_k has bounded derivatives and satisfies Assumption 2.3. For each k, the function ϕ_k is supported on $[(k-1)/m, k/m]$, and the integral of ϕ_k is 0, since

$$\begin{aligned}
\int \phi_k(x)\, dx &= Lh_1^{\beta_1} \int K\left(\frac{x - x_k}{h_1}\right) dx \\
&= Lh_1^{\beta_1} \int_{x_k - \frac{h}{2}}^{x_k + \frac{h}{2}} K\left(\frac{x - x_k}{h_1}\right) dx \qquad [\text{Since } K(z) > 0 \text{ when } |z| \geq 1/2] \\
&= Lh_1^{\beta_1+1} \int_{-\frac{1}{2}}^{\frac{1}{2}} K(z)\, dz = 0\,.
\end{aligned}$$

For $j \neq k$, ϕ_j and ϕ_k have disjoint support, yielding the $L_2([0,1])$ inner product between any ϕ_j and ϕ_k is 0. Set $M_1 := dm_1$ and $\Omega_1 := \{0,1\}^{M_1}$. For each $\boldsymbol{\omega}_1 \in \Omega_1$, define its entries by $\{\boldsymbol{\omega}_{1,jk}\}_{1\leq j\leq d, 1\leq k\leq m_1}$. Define functions $f_{\omega_1}^{(1)}$ as:

$$f_{\boldsymbol{\omega}_1}^{(1)}(X) := \sum_{j=1}^{d} \sum_{k=1}^{m} \boldsymbol{\omega}_{1,jk} \phi_k(X_j)\,,$$

which will contribute to the one dimensional alternatives.

Alternatives for two-dimensional components: The construction for two-dimensional alternatives are similar to that for one-dimensional components. As before, divide $[0,1]^2$ into m_2^2 rectangles $[(k-1)/m_2, k/m_2] \times [(l-1)/m_2, l/m_2]$, $1 \leq k, l \leq m_2$. Set $x_k := (k-0.5)/m_2$. Using K same as before, define ϕ_{kl}, $1 \leq k, l \leq m_2$ as

$$\phi_{kl}(x,y) := Lh_2^{\beta_2} K\left(\frac{x - x_k}{h}\right) K\left(\frac{y - x_l}{h}\right) \tag{A.6}$$

Observe that, each ϕ_{kl} is supported on $[(k-1)/m_2, k/m_2] \times [(l-1)/m_2, l/m_2]$. and consequently $L_2([0,1]^2)$ inner-product of any two ϕ_{k_1,l_1} and ϕ_{k_2,l_2} is 0 if $(k_1,l_1) \neq (k_2,l_2)$. Also the marginals of ϕ_{kl} are 0 as $\int K((x-x_k)/h)\, dx = 0$. Set $M_2 := d(d-1)m_2^2/2$ and $\Omega_2 := \{0,1\}^{M_2}$. For each $\boldsymbol{\omega}_2 \in \Omega_2$ we write it in a tensor form $\boldsymbol{\omega}_{2,ijkl}$ where $1 \leq i < j \leq d$ and $1 \leq k \leq l \leq m_2$ and define $f^{(2)}_{\boldsymbol{\omega}_2}$ as:

$$f^{(2)}_{\boldsymbol{\omega}_2}(X) = \sum_{i<j} \sum_{k,l} \boldsymbol{\omega}_{2,ijkl} \phi_{kl}(X_i, X_j)\,.$$

These functions will be the two-dimensional component of our alternatives.

Final step: We now construct our alternatives by combining the one-dimensional and two-dimensional components as constructed above. We first set the true mean function $f_0 = 0$. Now we choose $S_1 \subset \Omega_1$ and $S_2 \subset \Omega_2$ carefully and then construct our alternatives as:

$$\mathcal{F}_S = \left\{ f_{\boldsymbol{\omega}} = f^{(1)}_{\boldsymbol{\omega}_1} + f^{(2)}_{\boldsymbol{\omega}_2},\ \boldsymbol{\omega}_1 \in S_1, \boldsymbol{\omega}_2 \in S_2 \right\}$$

To choose S_1 and S_2 we invoke Varshamov-Gilbert theorem, cf. (Tsybakov, 2004, Lemma 2.9) which we state here for the convenience of the reader:

PROPOSITION A.2 (Varshamov-Gilbert). *For any $M \geq 8$, there exists $S \subseteq \{0,1\}^M$ with $|S| \geq 2^{M/8}$, such that for any $\boldsymbol{\omega} \neq \boldsymbol{\omega}' \in S$, $\rho(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq M/8$, where ρ is the Hamming distance.*

By Proposition A.2, we can choose S_1 and S_2 such that for any $\boldsymbol{\omega}_1, \boldsymbol{\omega}_1' \in S_1$ we have $\rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}_1') \geq M_1/8$ and for any $\boldsymbol{\omega}_2, \boldsymbol{\omega}_2' \in S_2$, $\rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}_2') \geq M_2/8$. Observe that for any $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) \neq \boldsymbol{\omega}' = (\boldsymbol{\omega}_1', \boldsymbol{\omega}_2')$ we have:

$$\begin{aligned}
d^2\left(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}\right) &= \int_{[0,1]^d} \left(f_{\boldsymbol{\omega}}(\mathbf{x}) - f_{\boldsymbol{\omega}'}(\mathbf{x})\right)^2 d\mathbf{x} \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) + \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right)^2 dx \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) \right)^2 d\mathbf{x} + \left(\sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right)^2 d\mathbf{x} \\
&\qquad + 2 \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) \right) \left(\sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right) d\mathbf{x} \\
&\triangleq T_1 + T_2 + 2T_3\,.
\end{aligned}$$

We now analyze each T_i separately.

$$\begin{aligned}
T_1 &= \int_{[0,1]^d} \left(\sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk}) \phi_k(X_j) \right)^2 d\mathbf{x} \\
&= \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 \int_0^1 \phi_k^2(X_j)\, dX_j
\end{aligned}$$

$$= L^2 h_1^{2\beta_1} \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 \int_0^1 K^2 \left(\frac{X_j - x_k}{h_1} \right) dX_j$$

$$= L^2 h_1^{2\beta_1+1} \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} K^2 (z) \, dz$$

$$= L^2 h_1^{2\beta_1+1} \|K\|_2^2 \sum_{j=1}^{d} \sum_{k=1}^{m} (\boldsymbol{\omega}_{1,jk} - \boldsymbol{\omega}'_{1,jk})^2 = L^2 h_1^{2\beta_1+1} \|K\|_2^2 \, \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}'_1)$$

For T_2:

$$T_2 = \int_{[0,1]^d} \left(\sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl}) \phi_{kl}(X_i, X_j) \right)^2 d\mathbf{x}$$

$$= \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 \int_{[0,1]^2} \phi_{kl}^2(X_i, X_j) \, dX_i \, dX_j$$

$$= L^2 h_2^{2\beta_2} \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 \int_{[0,1]^2} K^2 \left(\frac{X_i - x_j}{h_2} \right) K^2 \left(\frac{X_j - x_k}{h_2} \right) dX_i \, dX_j$$

$$= L^2 h_2^{2\beta_2+2} \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} K^2(z) \, dz \right)^2$$

$$= L^2 h_2^{2\beta_2+2} \|K\|_2^4 \sum_{i<j} \sum_{k,l} (\boldsymbol{\omega}_{2,ijkl} - \boldsymbol{\omega}'_{2,ijkl})^2 = L^2 h_2^{4\beta_2+2} \|K\|_2^4 \rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}'_2) \, .$$

Finally we claim that $T_3 = 0$. This follows from the fact that if i, j, k all are distinct, then for any $1 \le v_1 \le m_1$ and $1 \le v_2, v_3 \le m_2$:

$$\int_{[0,1]^3} \phi_{v_1}(X_i) \phi_{v_2,v_3}(X_j, X_k) \, dX_i \, dX_j \, dX_k$$

$$= \int_{[0,1]} \phi_{v_1}(X_i) \, dX_i \int_{[0,1]^2} \phi_{v_2,v_3}(X_j, X_k) \, dX_j \, dX_k = 0$$

as the integrals of ϕ_{v_1} and ϕ_{v_2,v_3} are 0 by construction. When they are not all distinct, suppose $i = j \neq k$. Then:

$$\int_{[0,1]^2} \phi_{v_1}(X_i) \phi_{v_2,v_3}(X_i, X_k) \, dX_i \, dX_k$$

$$= \int_{[0,1]} \phi_{v_1}(X_i) \left(\int_{[0,1]} \phi_{v_2,v_3}(X_i, X_k) \, dX_k \right) dX_i = 0$$

as marginals of ϕ_{v_2,v_3} is 0 by our construction. Therefore $T_3 = 0$. Combining the expression of T_i's, we have:

$$d^2(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) = L^2 h_1^{2\beta_1+1} \|K\|_2^4 \, \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}'_1) + L^2 h_2^{2\beta_2+2} \|K\|_2^4 \, \rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}'_2) \tag{A.7}$$

where ρ is the Hamming distance. From Proposition A.2, we have:

$$d^2(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) \ge L_1^2 h_1^{2\beta_1+1} \|K\|_2^2 \, \frac{dm_1}{8} + L_2^2 h_2^{2\beta_2+2} \|K\|_2^4 \, \frac{d(d-1)m_2^2}{16}$$

$$= \frac{L_1^2\|K\|_2^2}{8} d m_1^{-2\beta_1} + \frac{L_2^2\|K\|_2^4}{8}\binom{d}{2} m_2^{-2\beta_2} := 4\delta^2 . \tag{A.8}$$

Here the second equality follows from the fact that $h_1 = m_1^{-1}$ and $h_2 = m_2^{-1}$ and the third inequality follows from the definition of m_1, m_2. Applying Fano's inequality (Mukherjee et al., 2021, Proof of Theorem 2.18) on the collection $\mathcal{F}_S$, we obtain:

$$\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq \delta^2 \left(1 - \frac{\frac{n}{|S|^2}\sum_{\omega \neq \omega' \in S} KL\left(\mathbb{P}_\omega | \mathbb{P}_{\omega'}\right) + \log 2}{\log\left(|S| - 1\right)}\right) \tag{A.9}$$

As the error ϵ_i's are normal, we have:

$$\begin{aligned}
KL\left(\mathbb{P}_\omega | \mathbb{P}_{\omega'}\right) &= \frac{1}{2}\mathbb{E}\left[\left(f_\omega(X) - f_{\omega'}(X)\right)^2\right] = \frac{1}{2} d^2\left(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}\right) \\
&= L^2 h_1^{2\beta_1+1}\|K\|_2^2\, \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}_1') + L^2 h_2^{2\beta_2+2}\|K\|_2^4\, \rho(\boldsymbol{\omega}_2, \boldsymbol{\omega}_2') \\
&\leq L^2 h_1^{2\beta_1+1}\|K\|_2^2\, M_1 + L^2 h_2^{2\beta_2+2}\|K\|_2^4\, M_2 \\
&= L^2 h_1^{2\beta_1+1}\|K\|_2^2\, d m_1 + L^2 h_2^{2\beta_2+2}\|K\|_2^4 \binom{d}{2} m_2^2 \\
&= L^2\|K\|_2^2 d m_1^{-2\beta_1} + L^2\|K\|_2^4\binom{d}{2} m_2^{-2\beta_2} \leq 32\delta^2 .
\end{aligned}$$

Moreover, $|S| = |S_1 \times S_2| \geq 2^{(M_1+M_2)/8}$ by Proposition A.2. Plugging the bounds in (A.9), we have:

$$\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq \delta^2 \left(1 - 9\frac{32n\delta^2 + \log 2}{M_1 + M_2}\right) \tag{A.10}$$

Now from the definition of m_1, m_2 we have:

$$\begin{aligned}
n\delta^2 &= n\left(\frac{L_1^2\|K\|_2^2}{32} d m_1^{-2\beta_1} + \frac{L_2^2\|K\|_2^4}{32}\binom{d}{2} m_2^{-2\beta_2}\right) \\
&\leq Cn\left(\frac{L_1^2\|K\|_2^2}{32} d n^{-\frac{2\beta_1}{2\beta_1+1}} + \frac{L_2^2\|K\|_2^4}{32}\binom{d}{2} n^{-\frac{2\beta_2}{2\beta_2+2}}\right) \\
&= C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right)\left(d n^{\frac{1}{2\beta_1+1}} + \binom{d}{2} n^{\frac{2}{2\beta_2+2}}\right) \\
&\triangleq C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right)\psi_n .
\end{aligned}$$

On the other hand:

$$M_1 + M_2 = d m_1 + \binom{d}{2} m_2^2 \geq c\left(d n^{\frac{1}{2\beta_1+1}} + \binom{d}{2} n^{\frac{2}{2\beta_2+2}}\right) = c\psi_n .$$

Using these bounds in (A.10) yields:

$$\begin{aligned}
\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] &\geq \delta^2 \left(1 - 9\frac{C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right)\psi_n + \log 2}{c\psi_n}\right) \\
&= \delta^2 \left(1 - 9\frac{C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right) + \frac{\log 2}{\psi_n}}{c}\right)
\end{aligned}$$

The constants C, c depends on c_1, c_2 via the definition of m_1, m_2. Choosing them appropriately and using the fact $\psi_n \uparrow \infty$ as $n \uparrow \infty$, we can make

$$9\frac{C\left(\frac{L_1^2\|K\|_2^2}{32} \vee \frac{L_2^2\|K\|_2^4}{32}\right) + \frac{\log 2}{\psi_n}}{c} \leq \frac{1}{2}.$$

which implies:

$$\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \geq \delta^2$$
$$\geq \left(\frac{L_1^2\|K\|_2^2}{8} \wedge \frac{L_2^2\|K\|_2^4}{8}\right)\left(dn^{\frac{-2\beta_1}{2\beta_1+1}} + \binom{d}{2} n^{\frac{-2\beta_2}{2\beta_2+2}}\right).$$

This completes the proof.

A.3. Proof of Theorem 3.2 and 3.5. Recall the for any $\phi = \sum_j \phi_j + \sum_{k<l} \phi_{kl}$, we define $\|\phi\|_{n,1} := \sum_j \|\phi_j\|_n + \sum_{j<k} \|\phi_{jk}\|_n$ and $\|\phi\|_n^2 = \sum_{i=1}^n \phi^2(X_i)/n$. Further define $\|\phi\|_{n,lin} := \sum_j \|\phi_j\|_n$ and $\|\phi\|_{n,quad} := \sum_{k<l} \|\phi_{kl}\|_n$. From Theorem 2.9, there exists $\{\phi_j^\star\}_{j\in S_1}, \{\phi_{kl}^\star\}_{(j,k)\in S_2}$, $j \in S_1, (j,k) \in S_2$, such that, $\phi_j^\star \in \mathcal{F}_{NN,1}$, $\phi_{jk}^\star \in \mathcal{F}_{NN,2}$ and

$$\|f_{0,j} - \phi_j^\star\|_\infty \leq C_1(N_1L_1)^{-2\beta_1} \; \forall \; j \in S_1, \tag{A.11}$$

$$\|f_{0,kl} - \phi_{kl}^\star\|_\infty \leq C_2(N_2L_2)^{-\beta_2} \; \forall \; (k<l) \in S_2. \tag{A.12}$$

and furthermore, $\phi_j^\star$ has integral 0 and $\phi_{kl}^\star$ has 0 marginals. For the simplicity of the rest of the proof, take L_i to be of the constant order. Therefore the architecture we use here is the following: for the fitting the univariate function we use DNNs from the class $\mathcal{F}(1, c_1N_1\log N_1, c_2, 1)$ and for fitting bivariate functions, we use DNNs from the class $\mathcal{F}(1, c_1N_2\log N_2, c_3, 1)$. Therefore, defining $\phi^\star = \sum_{j\in S_1}\phi_j^\star + \sum_{(j,k)\in S_2}\phi_{jk}^\star$, we have by Theorem 2.7:

$$\mathbb{E}\left[(f_0(X) - \phi^\star(X))^2\right] \leq C_1\left(s_1(N_1)^{-4\beta_1} + s_2(N_2)^{-2\beta_2}\right)$$
$$\triangleq C_1(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2). \tag{A.13}$$

As $\hat{\phi}$ is the minimizer of the penalized loss function among the class of neural networks, it outperforms $\phi^\star$. Setting $pen(\phi) := \lambda_{n,1}\|\phi\|_{n,lin} + \lambda_{n,2}\|\phi\|_{n,quad}$, we have:

$$\begin{aligned}
&\frac{1}{2}\|Y - \hat{\phi}\|_n^2 + pen(\hat{\phi}) \leq \frac{1}{2}\|Y - \phi^\star\|_n^2 + pen(\phi^\star) \\
\implies &\frac{1}{2}\|\hat{\phi} - \phi^\star\|_n^2 + pen(\hat{\phi}) \leq \langle f_0 - \phi^\star, \hat{\phi} - \phi^\star\rangle_n + \langle \epsilon, \hat{\phi} - \phi^\star\rangle_n + pen(\phi^\star) \\
\implies &\frac{1}{2}\|\hat{\phi} - \phi^\star\|_n^2 + pen(\hat{\phi}) \leq \|f_0 - \phi^\star\|_n^2 + \frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \langle \epsilon, \hat{\phi} - \phi^\star\rangle_n + pen(\phi^\star) \\
\implies &\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + pen(\hat{\phi}) \leq \|f_0 - \phi^\star\|_n^2 + \langle \epsilon, \hat{\phi} - \phi^\star\rangle_n + pen(\phi^\star).
\end{aligned} \tag{A.14}$$

First we bound the empirical error $\|f_0 - \phi^\star\|_n^2$ in terms of its population counterpart $\|f_0 - \phi^\star\|_2^2$. Using Chebychev inequality, we have for $t > 0$:

$$\mathbb{P}\left(\left|\|f_0 - \phi^\star\|_n^2 - \|f_0 - \phi^\star\|_2^2\right| > t\right) \leq \frac{\mathbb{E}\left[(f_0 - \phi^\star)^4\right]}{nt^2} \tag{A.15}$$

Expanding the fourth moment, we obtain:

$$\mathbb{E}\left[(f_0-\phi^\star)^4\right] \le 8\left\{\mathbb{E}\left[\left(\sum_{j\in S_1}(f_{0,j}-\phi^\star_j)\right)^4\right]+\mathbb{E}\left[\left(\sum_{(j,k)\in S_2}(f_{0,jk}-\phi^\star_{jk})\right)^4\right]\right\}$$

$$\le 8p_{\max}\left\{\int_{[0,1]^d}\left(\sum_{j\in S_1}(f_{0,j}(x_j)-\phi^\star_j(x_j))\right)^4 d\mathbf{x}+\int_{[0,1]^d}\left(\sum_{(j,k)\in S_2}(f_{0,jk}(x_j,x_k)-\phi^\star_{jk}(x_j,x_k))\right)^4 d\mathbf{x}\right\}$$

$$\triangleq 8p_{\max}(T_1+T_2)\,.$$

Now we analyze T_1 and T_2 separately. For T_1:

$$T_1=\int_{[0,1]^d}\left(\sum_{j\in S_1}(f_{0,j}(x_j)-\phi^\star_j(x_j))\right)^4 d\mathbf{x}$$

$$=\sum_{j\in S_1}\int_0^1(f_{0,j}(x_j)-\phi^\star_j(x_j))^4\,dx_j+\sum_{j\neq k\in S_1}\int_0^1(f_{0,j}(x_j)-\phi^\star_j(x_j))^2\,dx_j\int_0^1(f_{0,j}(x_k)-\phi^\star_k(x_k))^2\,dx_k$$

$$\le C^4 s_1^2(N_1L_1)^{-8\beta}\,.$$

where the last line follows from the fact that $\|f_{0,j}-\phi^\star_j\|_\infty \le C(N_1L_1)^{-2\beta}$. Similar analysis for T_2 yields:

$$T_2=\int_{[0,1]^d}\left(\sum_{(j,k)\in S_2}(f_{0,jk}(x_j,x_k)-\phi^\star_{jk}(x_j,x_k))\right)^4 d\mathbf{x}\le C^4 s_2^2(N_2L_2)^{-4\beta}\,.$$

using $\|f_{0,jk}-\phi^\star_{jk}\|_\infty \le C(N_2L_2)^{-\beta}$, repeatedly and using Assumption 3.1. Using the bounds of T_1 of T_2, we have:

$$\mathbb{E}\left[(f_0-\phi^\star)^4\right] \le 8C^4p_{\max}\left(s_1^2(N_1L_1)^{-8\beta}+s_2^2(N_2L_2)^{-4\beta}\right)$$

$$\le 8C^4p_{\max}\left(s_1^2\rho_{n,1}^4+s_2^2\rho_{n,2}^4\right).$$

Hence, choosing $t_0=C^2\sqrt{\frac{8p_{\max}(s_1^2\rho_{n,1}^4+s_2^2\rho_{n,2}^4)\log n}{n}}$, we have from (A.15):

$$\mathbb{P}\left(\left|\|f_0-\phi^\star\|_n^2-\|f_0-\phi^\star\|_2^2\right|>t_0\right)\le\frac{1}{\log n} \tag{A.16}$$

Define the event $\Omega_{n,1}$ as:

$$\Omega_{n,1}=\left\{\left|\|f_0-\phi^\star\|_n^2-\|f_0-\phi^\star\|_2^2\right|\le t_0\right\}.$$

Noticing $t_0 \ll s_1\rho_{n,1}^2+s_2\rho_{n,2}^2$, by (A.13), we have on $\Omega_{n,1}$ that

$$\text{(A.14)} \implies \frac{1}{4}\|\hat\phi-\phi^\star\|_n^2+pen(\hat\phi)\le C_3(s_1\rho_{n,1}^2+s_2\rho_{n,2}^2)+\langle\epsilon,\hat\phi-\phi^\star\rangle_n+pen(\phi^\star)$$

In the next step, we bound the empirical error, i.e. the inner product between ϵ and $\hat\phi-\phi^\star$. First of all, note that, by triangle inequality:

$$\left|\langle\epsilon,\hat\phi-\phi^\star\rangle_n\right|\le\sum_{j=1}^d\left|\frac{1}{n}\sum_i\epsilon_i\left(\hat\phi_j(X_{ij})-\phi^\star_j(X_{ij})\right)\right|$$

$$+\sum_{k<l}\left|\frac{1}{n}\sum_i \epsilon_i\left(\hat{\phi}_{kl}(X_{ik},X_{il})-\phi^\star_{kl}(X_{ik},X_{il})\right)\right|$$

To bound each of the summands on the RHS, we first bound each term inside the summation. Towards that direction, we use Lemma 4 of Fan and Gu (2022), which says that if $\mathcal{G}_n$ is a class of functions with VC dimension V_n, then for any fixed $g_0 \in \mathcal{G}, \epsilon > 0, t > 0$, with probability $\geq 1 - \log(1/\epsilon)e^{-t}$:

$$\frac{1}{n}\left|\sum_{i=1}^n \epsilon_i(g(X_i)-g_0(X_i))\right| \leq C_4\left(\|g-g_0\|_n+\epsilon\right)\sqrt{\frac{V_{n,1}\log n}{n}+\frac{t}{n}}$$

for some universal constant c. Note that, $\phi_j \in \mathcal{F}_{NN}(1, c_1 N_1 \log N_1, c_2, c_3(N_1 \log N_1)^2, 1)$, whose VC dim $V_{n,1} \leq C_5 N_1^2 \log N_1)$ (see Lemma 2.8). We now apply this lemma to each of the component functions with $t = 2\log d$ and use a union bound to conclude:

(A.17)
$$\sum_{j=1}^d\left|\frac{1}{n}\sum_i \epsilon_i\left(\hat{\phi}_j(X_{ij})-\phi^\star_j(X_{ij})\right)\right| \leq C_4\sqrt{\frac{V_{n,1}\log p}{n}+\frac{2\log d}{n}}\left(\sum_{j=1}^d\|\hat{\phi}_j-\phi^\star_{n,j}\|_n+d\epsilon_1\right).$$

The above event occurs with probability $\geq 1-\log(1/\epsilon_1)e^{\log d - 2\log d} = 1-\log(1/\epsilon_1)/d$. A similar calculation for the bivariate components yields:

$$\sum_{k<l}\left|\frac{1}{n}\sum_i \epsilon_i\left(\hat{\phi}_{kl}(X_{ik},X_{il})-\phi^\star_{kl}(X_{ik},X_{il})\right)\right|$$

(A.18)
$$\leq C_5\sqrt{\frac{V_{n,2}\log n}{n}+\frac{3\log d}{n}}\left(\sum_{k<l}\|\hat{\phi}_{kl}-\phi^\star_{kl}\|_n+\frac{d(d-1)}{2}\epsilon_2\right),$$

with probability $\geq 1-\log(1/\epsilon_2)e^{2\log d-3\log d} = 1-\log(1/\epsilon_2)/d$. For the rest of the analysis, define the penalty parameters $\lambda_{n,1}, \lambda_{n,2}$ and ϵ_1, ϵ_2 as:

(A.19)
$$\lambda_{n,1} = 2C_4\sqrt{\frac{V_{n,1}\log p}{n}+\frac{2\log d}{n}}$$

(A.20)
$$\lambda_{n,2} = 2C_5\sqrt{\frac{V_{n,2}\log n}{n}+\frac{3\log d}{n}}$$

(A.21)
$$\epsilon_i = \frac{s_i\lambda_{n,i}}{\binom{d}{i}} \text{ for } i = 1,2.$$

Combining equation (A.17) and (A.18) we have:

$$\left|\langle\epsilon,\hat{\phi}-\phi^\star\rangle_n\right| \leq \frac{\lambda_{n,1}}{2}\sum_{j=1}^p\|\hat{\phi}_j-\phi^\star_{n,j}\|_{n,lin}+\frac{\lambda_{n,2}}{2}\sum_{k<l}\|\hat{\phi}_{kl}-\phi^\star_{kl}\|_{n,quad}$$

(A.22)
$$+\frac{s_1\lambda_{n,1}^2}{2}+\frac{s_2\lambda_{n,2}^2}{2}.$$

Define this event (A.22) to be $\Omega_{n,2}$ and we will later show that our choice of (N_1, N_2) ensures that $\mathbb{P}(\Omega_{n,2}) \to 1$. On the event $\Omega_{n,1} \cap \Omega_{n,2}$, we have:

$$\frac{1}{4}\|\hat{\phi}-\phi^\star\|_n^2+\lambda_{n,1}\|\hat{\phi}\|_{n,lin}+\lambda_{n,2}\|\hat{\phi}\|_{n,quad}$$

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{\lambda_{n,1}}{2}\sum_{j=1}^{p}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin}$$

$$+ \frac{\lambda_{n,2}}{2}\sum_{k<l}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad} + \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2} + \lambda_{n,1}\|\hat{\phi}\|_{n,lin} + \lambda_{n,2}\|\hat{\phi}\|_{n,quad}$$

Some simple algebra yields:

$$\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \frac{\lambda_{n,1}}{2}\|\hat{\phi}\|_{n,lin} + \frac{\lambda_{n,2}}{2}\|\hat{\phi}\|_{n,quad}$$

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{3\lambda_{n,1}}{2}\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin} + \frac{3\lambda_{n,2}}{2}\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}$$

(A.23)

$$+ \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2}\,.$$

If we don't assume any *Restricted Strong Convexity*(RSC) on the underlying function class (i.e. in our case DNNs with some pre-specified architecture), then we can use the fact $\|\hat{\phi}_j - \phi_{n,j}^\star\|_n \leq 2B$ for j and consequently we have:

$$\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \frac{\lambda_{n,1}}{2}\|\hat{\phi}\|_{n,lin} + \frac{\lambda_{n,2}}{2}\|\hat{\phi}\|_{n,quad}$$

(A.24)

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{3}{2}(s_1\lambda_{n,1} + s_2\lambda_{n,2}) + \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2}\,.$$

This completes the proof of Theorem 3.2. For the second part, we assume to have RSC condition Assumption (3.4). From (A.23), we obtain:

$$\frac{1}{4}\|\hat{\phi} - \phi^\star\|_n^2 + \frac{\lambda_{n,1}}{2}\|\hat{\phi}\|_{n,lin} + \frac{\lambda_{n,2}}{2}\|\hat{\phi}\|_{n,quad}$$

$$\leq C_3(s_1\rho_{n,1}^2 + s_2\rho_{n,2}^2) + \frac{3\lambda_{n,1}}{2}\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin} + \frac{3\lambda_{n,2}}{2}\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}$$

$$+ \frac{s_1\lambda_{n,1}^2}{2} + \frac{s_2\lambda_{n,2}^2}{2}$$

$$\leq C_3(s_1(\rho_{n,1}^2 + \lambda_{n,1}^2) + s_2(\rho_{n,2}^2 + \lambda_{n,2}^2)) + \frac{3}{2}\lambda_{n,1}\sqrt{s_1}\sqrt{\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin}^2}$$

$$+ \frac{3}{2}\lambda_{n,2}\sqrt{s_2}\sqrt{\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}^2}$$

$$\leq C_3(s_1(\rho_{n,1}^2 + \lambda_{n,1}^2) + s_2(\rho_{n,2}^2 + \lambda_{n,2}^2)) + \frac{9s_1\lambda_{n,1}^2}{2\kappa_1^2} + \frac{\kappa_1^2}{8}\sum_{j\in S_1}\|\hat{\phi}_j - \phi_{n,j}^\star\|_{n,lin}^2$$

$$+ \frac{9s_2\lambda_{n,2}^2}{2\kappa_2^2} + \frac{\kappa_2^2}{8}\sum_{k<l\in S_2}\|\hat{\phi}_{kl} - \phi_{kl}^\star\|_{n,quad}^2$$

$$\leq C_3(s_1(\rho_{n,1}^2 + \lambda_{n,1}^2) + s_2(\rho_{n,2}^2 + \lambda_{n,2}^2)) + \frac{9s_2\lambda_{n,2}^2}{2\kappa_2^2} + \frac{9s_1\lambda_{n,1}^2}{2\kappa_1^2} + \frac{1}{8}\|\hat{\phi} - \phi^\star\|_n^2$$

This implies:

$$\frac{1}{8}\|\hat{\phi}-\phi^\star\|_n^2+\frac{\lambda_{n,1}}{2}\sum_{j\in S_1^c}\|\hat{\phi}_j\|_n+\frac{\lambda_{n,2}}{2}\sum_{(j<k)\in S_2^c}\|\hat{\phi}_{jk}\|_n \quad \le C_6\left(s_1(\rho_{n,1}^2+\lambda_{n,1}^2)+s_2(\rho_{n,2}^2+\lambda_{n,2}^2)\right)$$

Now we select the optimal value of N_i's by balancing $\rho_{n,i}$ and $\lambda_{n,i}$. First take $i=1$. Recall that we have:

$$\rho_{n,1}^2=N_1^{-4\beta_1},\ \text{and}\ \lambda_{n,1}^2=4C_4^2\left(\frac{V_{n,1}\log n}{n}+\frac{2\log d}{n}\right)\le C_7\left(\frac{N_1^2\log^3 N_1\log n}{n}+\frac{\log d}{n}\right)$$

Choosing $N_1=\lfloor n^{\frac{1}{2(2\beta_1+1)}}\rfloor$ yields:

$$\rho_{n,1}^2\sim n^{-\frac{2\beta_1}{2\beta_1+1}}, \tag{A.25}$$

$$\lambda_{n,1}^2\sim n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}. \tag{A.26}$$

In particular we have:

$$\rho_{n,1}^2+\lambda_{n,1}^2\le C_8\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}\right).$$

Similar calculation for $i=2$ implies that choosing $N_2=\lfloor n^{\frac{1}{2(\beta_2+1)}}\rfloor$ we have:

$$\rho_{n,2}^2+\lambda_{n,2}^2\le C_9\left(n^{-\frac{\beta_2}{\beta_2+1}}\log^4 n+\frac{\log d}{n}\right).$$

Therefore, the above choice of penalty parameters yields:

$$\begin{aligned}\frac{1}{8}\|\hat{\phi}-\phi^\star\|_n^2+\frac{\lambda_{n,1}}{2}\sum_{j\in S_1^c}\|\hat{\phi}_j\|_n+\frac{\lambda_{n,2}}{2}\sum_{(j<k)\in S_2^c}\|\hat{\phi}_{jk}\|_n&\\ \le C_{10}\left(s_1\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\log^4 n+\frac{\log d}{n}\right)+s_2\left(n^{-\frac{\beta_2}{\beta_2+1}}\log^4 n+\frac{\log d}{n}\right)\right).&\end{aligned}$$

This completes the proof.

A.4. Proof of Theorem 3.8. For notational simplicty, define $r_n^2=\tilde{C}(s_1\lambda_{n,1}^2+s_2\lambda_{n,2}^2)$ for some constant $\tilde{C}$ mentioned explicitly later in the proof. First we show that the set $\hat{S}_i$ for $i\in\{1,2\}$, obtained by hard thresholding (Step 3 of Algorithm 1) satisfies $\hat{S}_i\supset S_i$ with high probability (whp). Recall that $\|f_j^0\|_2>r_n$ and $\|f_{jk}^0\|_2>r_n$ for all $j\in S_1$ and $(j<k)\in S_2$. Using Lu et al. (2021), there exists DNNs $\{\phi_j^\star\}_j$ and $\{\phi_{jk}^\star\}$ such that $\|\phi_j^\star\|_2>r_n/2$ and $\|\phi_{jk}^\star\|_2>r_n/2$ as the approximation error for each component is $<r_n/2$ by the definition of r_n for all $j\in S_1$ and $(j<k)\in S_2$. We will show first that $S_1\subseteq\hat{S}_1$ whp. The key idea is as follows: if $S_1\nsubseteq\hat{S}_1$, there exists $j\in S_1$ such that $j\notin\hat{S}_1$, i.e. $\|\hat{\phi}_j\|_n<C_2\lambda_{n,1}$. Since $\|\phi_j^\star\|_2\ge r_n/2$, we have $\|\phi_j^\star\|_n\ge r_n/4$ whp (which again follows from the fact that $nr_n^2\to\infty$, details later). If $\|\hat{\phi}_j\|_n<C_2\lambda_{n,1}$, then

$$C(s_1\lambda_{n,1}^2+s_2\lambda_{n,2}^2)\ge\frac{1}{8}\|\hat{\phi}-\phi^\star\|_n^2\ge\frac{1}{8}\|\hat{\phi}_j-\phi_j^\star\|_n^2\ge\frac{1}{8}\left(\frac{r_n}{4}-C_2\lambda_{n,1}\right)^2\ge\frac{r_n^2}{2^7}. \tag{A.27}$$

This yields contradiction as soon as $\tilde{C}^2\ge 2^7C$. Now we rigorize this intuition. Define the following events:

1. $\Omega_{n,1}=\{\sum_{j\in S_1}\|\hat{\phi}_j-\phi_j^\star\|_n\le C(s_1\lambda_{n,1}+s_2\lambda_{n,2})\}$.

2. $\Omega_{n,2} = \left\{ \left| \|\phi_j^\star\|_n - \|\phi_j^\star\|_2 \right| \le \frac{\|\phi_j^\star\|_2^2}{2} \ \ \forall \ \ j \in S_1 \right\}$.
3. $\Omega_{n,3} = \left\{ \left| \|\phi_{jk}^\star\|_n - \|\phi_{jk}^\star\|_2 \right| \le \frac{\|\phi_{jk}^\star\|_2^2}{2} \ \ \forall \ \ (j<k) \in S_2 \right\}$.

Using the proof of Theorem 2.9 $\mathbb{P}(\Omega_{n,1}) \to 1$ as $n \to \infty$. For $\Omega_{n,2}$ note that:

$$\begin{aligned}
\sum_{j \in S_1} \mathbb{P}\left(\left| \|\phi_j^\star\|_n^2 - \|\phi_j^\star\|_2^2 \right| \ge \frac{\|\phi_j^\star\|_2^2}{2} \right) &\le \sum_{j \in S_1} \frac{4\mathsf{var}\left(\left(\phi_j^\star(X) \right)^2 \right)}{n\|\phi_j^\star\|_2^4} \\
&\le \sum_{j \in S_1} \frac{4\mathbb{E}\left(\left(\phi_j^\star(X) \right)^4 \right)}{n\|\phi_j^\star\|_2^4} \\
&\le \sum_{j \in S_1} \frac{4B^2\mathbb{E}\left(\left(\phi_j^\star(X) \right)^2 \right)}{n\|\phi_j^\star\|_2^4} \qquad [\text{As } \|\phi_j^\star\|_\infty \le B] \\
&= \frac{4s_1B^2}{n\|\phi_j^\star\|_2^2} \le \frac{4s_1B^2}{nr_n^2} .
\end{aligned}$$

Now by our choice of r_n, it is immediate that $\frac{nr_n^2}{s_1} \gtrsim \lambda_{n,1}^2 \gtrsim \log n \to \infty$, using (3.7). This implies $\mathbb{P}(\Omega_{n,2}^c) \to 0$. Similar calculation yields that $\mathbb{P}(\Omega_{n,3}^c) \to 0$. Now, on the event $\Omega_{n,1} \cap \Omega_{n,2} \cap \Omega_{n,3}$, we have from (A.27) that $S_i \subseteq \hat{S}_i$ for $i = 1, 2$.

For the rest of the calculation, define the event $\mathcal{T} = \{S_i \subseteq \hat{S}_i, i = 1, 2\}$. We have shown $\mathbb{P}(\mathcal{T}) = 1 - o(1)$. For the rest of the analysis, we assume $\mathcal{T}$ happens. Define $|\hat{S}_1 \cap S_1^c| = \gamma_1$ and $|\hat{S}_2 \cap S_2^c| = \gamma_2$. Therefore $|\hat{S}_1| = s_1 + \gamma_1$ and $|\hat{S}_2| = s_2 + \gamma_2$. Note that the values of γ_1, γ_2 satisfies:

$$\begin{aligned}
2C_{10}\left(\gamma_1\lambda_{n,1}^2 + \gamma_2\lambda_{n,2}^2 \right) &\le \frac{\lambda_{n,1}}{2} \sum_{j \in S_1^c} \|\hat{\phi}_j\|_n + \frac{\lambda_{n,2}}{2} \sum_{(j<k) \in S_2^c} \|\hat{\phi}_{jk}\|_n \\
&\le 2C_{10}\left(s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2 \right)
\end{aligned}$$

i.e.

$$\gamma_1\lambda_{n,1}^2 + \gamma_2\lambda_{n,2}^2 \le s_1\lambda_{n,1}^2 + s_2\lambda_{n,2}^2 . \tag{A.28}$$

According to Algorithm 1, we use the second half of the data to estimate the mean function by restricting ourselves only on $\hat{S}_1$ and $\hat{S}_2$ and minimizing (unpenalized) squared error loss, i.e. our final estimate is:

$$\hat{f}^{\text{final}}(x) = \sum_{j \in \hat{S}_1} \hat{\phi}_j^{\text{final}}(x_j) + \sum_{(j<k) \in \hat{S}_2} \hat{\phi}_{jk}^{\text{final}}(x_j, x_k),$$

where the component functions are estimated as:

$$\{\hat{\phi}_j^{\text{final}}\}, \{\hat{\phi}_{jk}^{\text{final}}\} = \arg\min_{\phi_j, \phi_{jk}} \frac{1}{n} \sum_i \left(Y_i - \sum_{j \in \hat{S}_1} \phi_j(X_{ij}) - \sum_{(j<k) \in \hat{S}_2} \phi_{jk}(X_{ij}, X_{ik}) \right)^2 .$$

Rest of the proof is similar to that of Theorem 2.9. By (Fan and Gu, 2022, Theorem 4.8), we know that if f is an d-variate β smooth function, then, there exists neural network g with

width N and constant depth (where the constant depends on d, β) such that

$$\|f - g\|_\infty \leq cN^{-\frac{2\beta}{d}} .$$

for some constant $c = c(\beta, d) > 0$. Here we fit univariate and bivariate functions seprately. Suppose we fit the univariate components using neural networks of width N_1 (and constant depth c_1) and bivariate components using neural networks of width N_2 (and constant depth c_2). Then to fit the additive functions, we need $|\hat{S}_1|$ many such univariate components and $|\hat{S}_2|$ many bivariate components. Therefore total number of weights (no. of active parameters) required to fit such a mean function is $W = C_1(|\hat{S}_1|N_1^2 + |\hat{S}_2|N_2^2)$. From Bartlett et al. (2019), that VC-dim of such neural networks is $V_n \leq C_2 W \log W$ (as the depth is $O(1)$). The bias-variance decomposition yields:

$$\|\hat{f}^{\text{final}} - f_0\|_{L_2(P_X)}^2 \leq 2 \left(\underbrace{\|\phi^\star - f_0\|_{L_2(P_X)}^2}_{\text{bias}} + \underbrace{\|\hat{f}^{\text{final}} - \phi^\star\|_{L_2(P_X)}^2}_{\text{variance}} \right)$$

where $\phi^\star$ is the best approximator of f_0 among the class of neural network over which we are optimizing (i.e. sum of $|\hat{S}_1|$ many univariate networks with width N_1 and sum of $|\hat{S}_2|$ many bivariate components of width N_2). As $S_i \subseteq \hat{S}_i$ for $i \in \{1, 2\}$, we know from the proof of Theorem 3.5 (see equation (A.13)) that

$$\|\phi^\star - f_0\|_{L_2(P_X)}^2 \leq C_1 \left(s_1 N_1^{-4\beta_1} + s_2 N_2^{-2\beta_2} \right) .$$

Now to bound the variance term we use VC dimension techniques similar to the proof of Theorem 2.9. For any choice of component function, we can treat the overall sum $\sum_{j \in \hat{S}_1} \phi_j + \sum_{(j<k) \in S_2} \phi_{jk}$ as a function from the VC class with VC-dim $\leq C_2 W \log W$. Therefore, from the proof of Theorem 2.9, we have:

$$\begin{aligned}
\|\hat{f}^{\text{final}} - f_0\|_{L_2(P_X)}^2 &= O_p \left(\frac{V_n \log n}{n} + s_1 N_1^{-4\beta_1} + s_2 N_2^{-2\beta_2} \right) \\
&= O_p \left(\frac{W \log W \log n}{n} + s_1 N_1^{-4\beta_1} + s_2 N_2^{-2\beta_2} \right) \\
&= O_p \left(\frac{(s_1 + \gamma_1) N_1^2 + (s_2 + \gamma_2) N_2^2 \log \left((s_1 + \gamma_1) N_1^2 + (s_2 + \gamma_2) N_2^2 \right) \log n}{n} \right. \\
&\qquad \left. + s_1 N_1^{-4\beta_1} + s_2 N_2^{-2\beta_2} \right) \\
&= O_p \left((s_1 + \gamma_1)\lambda_{n,1}^2 + (s_2 + \gamma_2)\lambda_{n,2}^2 \right) = O_p \left(s_1 \lambda_{n,1}^2 + s_2 \lambda_{n,2}^2 \right) ,
\end{aligned}$$

where the second last line follows from the definition of $\lambda_{n,1}$ and $\lambda_{n,2}$ (see (3.7)) and the last line follows from (A.28). This completes the proof.

A.5. Proof of Theorem 3.9. Here, we extend our proof of Theorem 2.12 to the sparse interaction model. The main technical change here is to incorporate sparsity. For that, we need slightly different definitions for m_1, m_2 from (A.4). We define m_1 to be the solution of:

$$m_1^{-2\beta_1} = c_1 \left(n^{-\frac{2\beta_1}{2\beta_1+1}} \vee 8 \frac{\log \left(\frac{2d}{s_1} - 2 \right)}{n} \right) = \frac{c_1}{n} \left(n^{\frac{1}{2\beta_1+1}} \vee 8 \log \left(\frac{2d}{s_1} - 2 \right) \right) \tag{A.29}$$

and m_2 to be the solution of the equation:

(A.30)
$$m_2^{-2\beta_2} = c_2 \left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee 8\frac{\log\left(\frac{d(d-1)}{s_2}-2\right)}{n} \right) = \frac{c_2}{n}\left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right)$$

If m_1, m_2 are not integers, we will take the nearest integer. As this does not affect us asymptotically, we henceforth assume them to be exact solution for the simplicity of proof. The constant $c_1, c_2 > 0$ will be chosen at the end of the proof. Similar to (A.4), define $h_i := m_i^{-1}$, $i \in \{1,2\}$. Consider the set $\Omega_i = \{0,1\}^{m_i}$ for $i \in \{1,2\}$. By Proposition A.2, we can find $S_i \subset \Omega_i$ such that $|S_i| \geq 2^{m_i/8}$ and for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in S_i$, we have $\rho(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq m_i/8$, where ρ is the Hamming distance. As before define $\Omega = \Omega_1 \times \Omega_2$ and $S = S_1 \times S_2$. For any $\boldsymbol{\omega} \in \Omega$, we write it as $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ where $\boldsymbol{\omega}_i \in \Omega_i$. For one-dimensional components, given any $\boldsymbol{\omega}_1 \in \Omega_1$, we define a function $f^{(1)}_{\omega_1}$ as:

$$f^{(1)}_{\boldsymbol{\omega}_1}(x) = \sum_{k=1}^{m} \boldsymbol{\omega}_{1,k}\phi_k(x)\,,$$

where ϕ_k defined in (A.5). Define $\mathcal{F}_1 := \{f^{(1)}_{\boldsymbol{\omega}_1} : \boldsymbol{\omega}_1 \in S_1\}$ and $\Gamma_1 := |\mathcal{F}_1| \geq 2^{m_1/8}$. Enumerate the functions in $\mathcal{F}_1$ as $\mathcal{F}_1 = \{f^{(1)}_1, \ldots, f^{(1)}_{\Gamma_1}\}$. Note that for any $i \neq j$:

$$\begin{aligned}
&\int_0^1 \left(f^{(1)}_i(x) - f^{(1)}_j(x)\right)^2 dx \\
&= \sum_{k=1}^{m} (\boldsymbol{\omega}_{i,k} - \boldsymbol{\omega}_{j,k})^2 \int_0^1 \phi_k^2(x)\, dx \quad [\text{Since } \langle \phi_j, \phi_k\rangle_{L_2} = 0 \;\forall\; j \neq k] \\
&= L_1^2 h^{2\beta_1+1} \|K\|_2^2 \rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \qquad \text{(A.31)}
\end{aligned}$$

For the two-dimensional components, given any $\boldsymbol{\omega}_2 \in \Omega_2$, we define $f^{(2)}_{\boldsymbol{\omega}_2}$ as:

$$f^{(2)}_{\boldsymbol{\omega}_2}(x,y) = \sum_{k,l=1}^{m} \boldsymbol{\omega}_{2,k,l}\phi_{k,l}(x,y)\,,$$

where $\phi_{k,l}$ is defined by (A.6). Define $\mathcal{F}_2 := \{f^{(2)}_{\boldsymbol{\omega}_2} : \boldsymbol{\omega}_2 \in S_2\}$ and $\Gamma_2 := |\mathcal{F}_2| \geq 2^{m_2^2/8}$. Enumerate the functions in $\mathcal{F}_2$ as $\mathcal{F}_2 = \{f^{(2)}_1, \ldots, f^{(2)}_{\Gamma_2}\}$. For any $i \neq j$:

$$\begin{aligned}
&\int_{[0,1]^2} \left(f^{(2)}_i(x,y) - f^{(2)}_j(x,y)\right)^2 dx\, dy \\
&= \sum_{1\leq k,l\leq m_2} (\boldsymbol{\omega}_{i,k,l} - \boldsymbol{\omega}_{j,k,l})^2 \iint_0^1 \phi^2_{k,l}(x,y)\, dx \quad [\text{Since } \langle \phi_{j,k}, \phi_{l,m}\rangle_{L_2} = 0 \;\forall\; (j,k) \neq (l,m)] \\
&= L_2^2 h^{2\beta_2+2} \|K\|_4^2 \rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \qquad \text{(A.32)}
\end{aligned}$$

To construct our set of sparse alternatives, define sets U_1^* and U_2^* as:

$$U_i^* = \left\{ u \in \{0, 1, \ldots, \Gamma_1\}^{\binom{d}{2}} : \|u\|_0 = s_i \right\}, \qquad i = 1, 2.$$

It is immediate that $|U_i^*| = \binom{d}{s_i}\Gamma_i^{s_i}$, $i = 1, 2$. Choose $U_i \subset U_i^*$ such that for any $u, v \in U_i$, we have $\rho(u, v) \geq s_i/2$. By the proof of (Raskutti et al., 2012, Lemma 4), we have:

$$|U_i| \geq \frac{1}{2}\left(\frac{\binom{d}{i} - s_i}{s_i/2}\right)^{s_i/2} \Gamma_i^{s_i/2}, \qquad i = 1, 2.$$

For $i \leq 2$, $u_i \in U_i$, define functions $f_{u_i}^{(i)}$ as:

$$f_{u_1}^{(1)}(X) = \sum_{j=1}^{d} f_{u_{1,j}}^{(1)}(X_j), \qquad f_{u_2}^{(2)}(X) = \sum_{1\leq i<j\leq d} f_{u_{2,i,j}}^{(2)}(X_i, X_j)$$

where $f_0^{(i)} = 0$, else it is chosen from $\mathcal{F}_i$. Now, our set of alternatives are:

$$f(X) = f_{u_1}^{(1)}(X) + f_{u_2}^{(2)}(X) \;\; u_1 \in U_1, u_2 \in U_2\,.$$

Define $\mathcal{F}^{\text{sparse}}$ to be collection of all such functions f. Let $M = |\mathcal{F}^{\text{sparse}}|$. Pick any two $f, f' \in \mathcal{F}^{\text{sparse}}$. We have:

$$\begin{aligned}
d^2(f, f') &= \int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j) + \sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX \\
&= \int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)^2 dX + \int_{[0,1]^d} \left(\sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX \\
&\qquad + 2\int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)\left(\sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right) dX \\
&\triangleq T_1 + T_2 + 2T_3
\end{aligned}$$

We now analyze each T_i separately.

$$\begin{aligned}
T_1 &= \int_{[0,1]^d} \left(\sum_{j=1}^{d}(f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)^2 dX \\
&= \sum_{j=1}^{d}\int_0^1 \left((f_{u_{1,j}}^{(1)} - f_{u'_{1,j}}^{(1)})(X_j)\right)^2 dX_j \quad \left[\text{Since} \int f_{u_1,j}^{(1)}(X_j)\, dX_j = 0\right] \\
&= L_1^2 h_1^{2\beta_1+1}\|K\|_2^2 \sum_{j=1}^{d} \rho(\boldsymbol{\omega}_{u_{1,j}}, \boldsymbol{\omega}_{u'_{1,j}})\, \mathbb{1}_{u_{1,j}\neq u'_{1,j}} \quad \text{[From (A.31)]}
\end{aligned}$$

$$\begin{aligned}
T_2 &= \int_{[0,1]^d} \left(\sum_{1\leq i<j\leq d}(f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX \\
&= \sum_{1\leq i<j\leq d}\int_{[0,1]^2} \left((f_{u_{2,i,j}}^{(2)} - f_{u'_{2,i,j}}^{(2)})(X_i, X_j)\right)^2 dX_i\, dX_j \quad [\text{ As marginals of } f^{(2)} \text{ are } 0] \\
&= L_2^2 h_2^{2\beta_2+2} \sum_{1\leq i<j\leq d} \rho(\boldsymbol{\omega}_{u_{2,i,j}}, \boldsymbol{\omega}_{u'_{2,i,j}})\, \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \quad \text{[From (A.32)]}
\end{aligned}$$

Furthermore, $T_3 = 0$ as the marginals of $f^{(2)}$ are 0. Combining the bounds of T_i's, we obtain:

$$
\begin{aligned}
d^2(f,f') &= L_1^2 h_1^{2\beta_1+1}\|K\|_2^2 \sum_{j=1}^{d} \rho(\boldsymbol{\omega}_{u_{1,j}}, \boldsymbol{\omega}_{u'_{1,j}})\, \mathbb{1}_{u_{1,j}\neq u'_{1,j}} \\
&\qquad + L_2^2 h_2^{2\beta_2+2}\|K\|_2^4 \sum_{1\le i<j\le d} \rho(\boldsymbol{\omega}_{u_{2,i,j}}, \boldsymbol{\omega}_{u'_{2,i,j}})\, \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \\
&\ge \frac{L_1^2\|K\|_2^2}{8} h_1^{2\beta_1+1} m_1 \sum_{j=1}^{d} \mathbb{1}_{u_{1,j}\neq u'_{1,j}} + \frac{L_2^2\|K\|_2^4}{8} h_2^{2\beta_2+2} m_2^2 \sum_{1\le i<j\le d} \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \\
&\ge \frac{L_1^2\|K\|_2^2}{16} h_1^{2\beta_1+1} m_1 s_1 + \frac{L_2^2\|K\|_2^4}{16} h_2^{2\beta_2+2} m_2^2 s_2 \\
&\ge \frac{L_1^2\|K\|_2^2}{16} m_1^{-2\beta_1} s_1 + \frac{L_2^2\|K\|_2^4}{16} m_2^{-2\beta_2} s_2 := 4\delta^2 . \qquad \text{(A.33)}
\end{aligned}
$$

Similarly we can obtain an upper bound in terms of δ:

$$
\begin{aligned}
d^2(f,f') &= L_1^2 h_1^{2\beta_1+1}\|K\|_2^2 \sum_{j=1}^{d} \rho(\boldsymbol{\omega}_{u_{1,j}}, \boldsymbol{\omega}_{u'_{1,j}})\, \mathbb{1}_{u_{1,j}\neq u'_{1,j}} \\
&\qquad + L_2^2 h_2^{2\beta_2+2}\|K\|_2^4 \sum_{1\le i<j\le d} \rho(\boldsymbol{\omega}_{u_{2,i,j}}, \boldsymbol{\omega}_{u'_{2,i,j}})\, \mathbb{1}_{u_{2,i,j}\neq u'_{2,i,j}} \\
&\le 2L_1^2\|K\|_2^2 s_1 m_1 h_1^{2\beta_1+1} + 2L_2^2\|K\|_2^4 s_2 m_2^2 h_2^{2\beta_2+2} \qquad \text{(A.34)} \\
&= 2L_1^2\|K\|_2^2 s_1 m_1^{-2\beta_1} + 2L_2^2\|K\|_2^4 s_2 m_2^{-2\beta_2} = 128\delta^2 . \qquad \text{(A.35)}
\end{aligned}
$$

Combining the bounds obtained in equation (A.33) and equation (A.34) we have that for any two $f, f' \in \mathcal{F}^{\text{sparse}}$:

$$
4\delta^2 \le d^2(f,f') \le 128\delta^2 . \qquad \text{(A.36)}
$$

which further implies that $\mathcal{F}^{\text{sparse}}$ is a 2δ packing set of the set of all feasible functions. An application of Fano's inequality (Mukherjee et al., 2021, Proof of Theorem 2.18) yields:

$$
\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \ge \delta^2 \left(1 - \frac{\frac{n}{M^2}\sum_{f,f'\in\mathcal{F}^{\text{sparse}}} KL\left(\mathbb{P}_f|\mathbb{P}_{f'}\right) + \log 2}{\log\left(M-1\right)}\right) \qquad \text{(A.37)}
$$

Now as the errors are normally distributed:

$$
KL\left(\mathbb{P}_f|\mathbb{P}_{f'}\right) = \mathbb{E}\left[\left(f(X) - f'(X)\right)^2\right] = d^2(f,f')
$$

Using the bounds of equation (A.36), we have $KL\left(\mathbb{P}_f|\mathbb{P}_{f'}\right) \le 128\delta^2$ for all f, f'. Therefore, we obtain from equation (A.37):

$$
\begin{aligned}
\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] &\ge \delta^2 \left(1 - \frac{128n\delta^2 + \log 2}{\log\left(M-1\right)}\right) \\
&\ge \delta^2 \left(1 - \frac{128n\delta^2 + \log 2}{2\log M}\right), \qquad \text{(A.38)}
\end{aligned}
$$

where $M = |U_1||U_2|$. The rest of the proof is to balance $n\delta^2$ and $\log M$ to say we can say $\frac{128n\delta^2+\log 2}{2\log M} \le c^*$ for some $0 < c^* < 1$. For that it is enough to show that $n\delta^2$ and $\log M$ has

same order as we can then chosen the constants carefully so that the ratio is within $(0,1)$. From the definition of δ, we have:

$$\begin{aligned} n\delta^2 &= \frac{L_1^2\|K\|_2^2}{64} n m_1^{-2\beta_1} s_1 + \frac{L_2^2\|K\|_2^4}{64} n m_2^{-2\beta_2} s_2 \\ &= \frac{L_1^2\|K\|_2^2}{4} n m_1^{-2\beta_1} \frac{s_1}{16} + \frac{L_2^2\|K\|_2^4}{4} n m_2^{-2\beta_2} \frac{s_2}{16} \\ &= \frac{L_1^2 c_1\|K\|_2^2}{4} \frac{s_1}{16} \left[n^{\frac{1}{2\beta_1+1}} \vee 8\log\left(\frac{2d}{s_1} - 2\right)\right] \\ &\qquad + \frac{L_2^2 c_2\|K\|_2^4}{4} \frac{s_2}{16} \left[(n^{\frac{1}{2\beta_2+2}} \vee 8\log\left(\frac{d(d-1)}{s_2} - 2\right)\right] \end{aligned} \tag{A.39}$$

On the other hand,by our construction, we have:

$$\begin{aligned} \log M &= \log|U_1| + \log|U_2| \\ &\geq 2\log\frac{1}{2} + \frac{s_1}{2}\log\left(\frac{d-s_1}{s_1/2}\right) + \frac{s_1}{2}\log\Gamma_1 + \frac{s_2}{2}\log\left(\frac{d(d-1)/2 - s_2}{s_2/2}\right) + \frac{s_2}{2}\log\Gamma_2 \\ &= 2\log\frac{1}{2} + \frac{s_1}{2}\log\left(\frac{2d}{s_1} - 2\right) + \frac{s_1}{2}\log\Gamma_1 + \frac{s_2}{2}\log\left(\frac{d(d-1)}{s_2} - 2\right) + \frac{s_2}{2}\log\Gamma_2 \\ &= 2\log\frac{1}{2} + \frac{s_1}{2}\log\left(\frac{2d}{s_1} - 2\right) + \frac{s_1 m_1}{16} + \frac{s_2}{2}\log\left(\frac{d(d-1)}{s_2} - 2\right) + \frac{s_2 m_2^2}{16} \\ &= 2\log\frac{1}{2} + \frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1} - 2\right)\right] + \frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2} - 2\right)\right] \end{aligned}$$

$$\triangleq 2\log\frac{1}{2} + a_n + b_n\,. \tag{A.40}$$

We now consider four cases:

Case 1: $n^{1/(2\beta_1+1)} \geq 8\log\left(\frac{2d}{s_1} - 2\right)$ **and** $n^{\frac{1}{2\beta_2+2}} \geq 8\log\left(\frac{d(d-1)}{s_2} - 2\right)$

Using (A.29), (A.30), we have $m_1 = n^{1/(2\beta_1+1)}$ and $m_2 = n^{1/(2\beta_2+2)}$. Therefore, equations (A.39) and (A.40) are simplified to:

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1\|K\|_2^2}{4}\frac{s_1}{16} n^{\frac{1}{2\beta_1+1}} + \frac{L_2^2 c_2\|K\|_2^4}{4}\frac{s_2}{16} n^{-\frac{2\beta_2}{2\beta_2+2}} \\ &\leq \frac{L_1^2 c_1\|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1} - 2\right)\right] + \frac{L_2^2 c_2\|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2} - 2\right)\right] \\ &\leq \frac{L_1^2 c_1\|K\|_2^2}{4} a_n + \frac{L_2^2 c_2\|K\|_2^4}{4} b_n \end{aligned}$$

Case 2: $n^{1/(2\beta_1+1)} \geq 8\log\left(\frac{2d}{s_1} - 2\right)$ **and** $n^{\frac{1}{2\beta_2+2}} < 8\log\left(\frac{d(d-1)}{s_2} - 2\right)$

Here, $m_1 = n^{1/(2\beta_1+1)}$ and $m_2 = \left(8\log\left(\frac{d(d-1)}{s_2}-2\right)/n\right)^{-1/2\beta_2}$ yielding

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16} n^{\frac{1}{2\beta_1+1}} + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16} 8\log\left(\frac{d(d-1)}{s_2}-2\right) \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1}-2\right)\right] + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right] \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n \end{aligned}$$

Case 3: $n^{1/(2\beta_1+1)} < 8\log\left(\frac{2d}{s_1}-2\right)$ **and** $n^{\frac{1}{2\beta_2+2}} \ge 8\log\left(\frac{d(d-1)}{s_2}-2\right)$

Here $m_1 = \left(8\log\left(\frac{2d}{s_1}-2\right)/n\right)^{-1/2\beta_1}$ and $m_2 = n^{\frac{1}{2\beta_2+2}}$ yielding

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16} 8\log\left(\frac{2d}{s_1}-2\right) + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16} n^{\frac{2}{2\beta_2+2}} \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1}-2\right)\right] + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right] \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n. \end{aligned}$$

Case 4: $n^{1/(2\beta_1+1)} < 8\log\left(\frac{2d}{s_1}-2\right)$ **and** $n^{\frac{1}{2\beta_2+2}} < 8\log\left(\frac{d(d-1)}{s_2}-2\right)$

Here $m_1 = \left(8\log\left(\frac{2d}{s_1}-2\right)/n\right)^{-1/2\beta_1}$ and $m_2 = \left(8\log\left(\frac{d(d-1)}{s_2}-2\right)/n\right)^{-1/2\beta_2}$ yielding

$$\begin{aligned} n\delta^2 &= \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16} 8\log\left(\frac{2d}{s_1}-2\right) + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16} 8\log\left(\frac{d(d-1)}{s_2}-2\right) \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4}\frac{s_1}{16}\left[m_1 + 8\log\left(\frac{2d}{s_1}-2\right)\right] + \frac{L_2^2 c_2 \|K\|_2^4}{4}\frac{s_2}{16}\left[m_2^2 + 8\log\left(\frac{d(d-1)}{s_2}-2\right)\right] \\ &\le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n. \end{aligned}$$

Hence, in all four cases, the numerator and the denominator has same order. Hence,

$$n\delta^2 \le \frac{L_1^2 c_1 \|K\|_2^2}{4} a_n + \frac{L_2^2 c_2 \|K\|_2^4}{4} b_n\,.$$

Putting this in equation (A.38) we obtain:

$$\begin{aligned} \inf_{\hat f}\sup_{f, P_X} \mathbb{E}_{P_X}[(\hat f(X) - f(X))^2] &\ge \delta^2\left(1 - \frac{32\left(L_1^2 c_1 \|K\|_2^2\, a_n + L_2^2 c_2 \|K\|_2^4\, b_n\right) + \log 2}{2(2\log\frac{1}{2} + a_n + b_n)}\right) \\ &\ge \delta^2\left(1 - \frac{33\left(L_1^2 c_1 \|K\|_2^2\, a_n + L_2^2 c_2 \|K\|_2^4\, b_n\right)}{3(a_n + b_n)}\right) \end{aligned}$$

$$\geq \delta^2 \left(1 - \frac{11\left(L_1^2 c_1 \|K\|_2^2 \, a_n + L_2^2 c_2 \|K\|_2^4 \, b_n\right)}{(a_n + b_n)}\right).$$

(A.41)

Now choose c_1, c_2 such that $(L_1^2 c_1 \|K\|_2^2) \vee (L_2^2 c_2 \|K\|_2^4) \leq 1/22$. Then we have:

$$\begin{aligned}
&\inf_{\hat{f}} \sup_{f, P_X} \mathbb{E}_{P_X}[(\hat{f}(X) - f(X))^2] \\
&\geq \frac{1}{2}\delta^2 \geq \frac{L_1^2 \|K\|_2^2}{128} m_1^{-2\beta_1} s_1 + \frac{L_2^2 \|K\|_2^4}{128} m_2^{-2\beta_2} s_2 \\
&= \frac{L_1^2 \|K\|_2^2 c_1}{128} s_1 \left(n^{-\frac{2\beta_1}{2\beta_1+1}} \vee 8 \frac{\log\left(\frac{2d}{s_1} - 2\right)}{n}\right) + \frac{L_2^2 \|K\|_2^4 c_2}{128} s_2 \left(n^{-\frac{2\beta_2}{2\beta_2+2}} \vee 8 \frac{\log\left(\frac{d(d-1)}{s_2} - 2\right)}{n}\right).
\end{aligned}$$

This completes the proof.