MSc Mathematics

Track: Probability and Statistics

Master thesis

Asymptotics of nonparametric regression using sparse neural networks with ReLU activation

by

Quoc Thanh Don Vu

November 28, 2022

Supervisor: dr. Eduard Belitser

Second reader: dr. Paulo Serra

Department of Mathematics

Faculty of Science

VU UNIVERSITY AMSTERDAM

Abstract

This thesis studies estimation of a nonparametric regression model by estimators based on sparsely connected deep neural networks with ReLU activation. We consider an earlier article on this topic and discuss its results in detail. These results show that such estimators can achieve good rates of convergence for regression functions with certain composition structures. In some cases, these rates can even be shown to be minimax up to $\log n$ factors. We discuss some minor points in the proofs of these results that were unclear to us, and we propose ways to modify the proofs to solve these problems with only small changes to the results. We also show that the results can be generalized to a setting with Markov chain regressors and subgaussian errors. In particular, our generalization allows for conditional and unconditional heteroskedasticity, and for dependence between the regressors and errors of different observations. The results discussed in this thesis might be a step towards a better understanding of why neural networks often perform well in practice, and in what situations they might do so.

Keywords: neural networks, deep learning, asymptotics, statistics, machine learning, nonparametric regression, sieve estimation, sparsity, Markov chains, time series.

Title: Asymptotics of nonparametric regression using sparse neural networks with ReLU activation

Author: Quoc Thanh Don Vu, q.t.d.vu@student.vu.nl, 2579602

Supervisor: dr. Eduard Belitser

Second reader: dr. Paulo Serra

Date: November 28, 2022

Department of Mathematics

Vrije Universiteit Amsterdam

De Boelelaan 1111, 1081 HV Amsterdam

`http://www.math.vu.nl/`

Popular summary

The world is full of situations in which you might want to predict some quantity Y based on some information $\boldsymbol{X}$. In many cases, you will want to do so based on examples you might have seen earlier. This is, very broadly speaking, the problem of regression that we consider in this thesis. More formally, let us suppose that a random vector $\boldsymbol{X}$ in $[0,1]^d$ and random variable Y in $\mathbb{R}$ are related by the equation $Y = f_0(\boldsymbol{X}) + \epsilon$, where f_0 is some unknown function and ϵ is some random error term with mean zero. In that case, we might want to predict Y by $f_0(\boldsymbol{X})$. Hence, we want to figure out what the regression function f_0 is. To be able to do so, we need data. We assume that we observe a sample $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ of independent observations that also come from the same process. That is, we also have $Y_i = f_0(\boldsymbol{X}_i) + \epsilon_i$ for every i. Based on this sample, we might then try make an educated guess about what f_0 is. This is called estimation, and we hope that our estimator of f_0 will converge quickly to f_0 as $n \to \infty$. Recently, estimation using neural networks has become popular. That is, we pick a neural network function as our educated guess of what f_0 might be. Neural networks are functions that can be visualized as a series of layers of nodes. Each node takes some input from the previous layer, transforms this, and passes the output to nodes in the next layer. The first layer is the input of the function, and the final layer is the output. Neural networks are a very flexible class of functions and they have shown impressive results for regression problems in practice. However, there is still a lack of statistical theory to explain why they are able to perform so well.

In this thesis, we consider a recent article that proves that regression functions with certain composition structures can be estimated well using certain neural networks. In some cases, it can even be shown that these estimators almost reach the best possible rate of convergence. We also prove a generalization of the main result of this article for situations in which the different observations in the sample might be dependent, rather than independent. This is relevant in many situations in practice, such as with time series data, or spatial data. The results discussed in this thesis can be seen as a step towards a better understanding of why neural networks are so effective in practice and when it might be a good idea to use them.

Contents

1 Introduction

Neural networks (NNs) have been celebrated for their impressive performance in regression problems. Examples of applications include colorizing black and white photographs (Iizuka et al., 2016), prediction of stock prices (Thakkar and Chaudhari, 2021) and prediction of traffic (Lv et al., 2014). The popularity of NNs in practice has also led to a rise in papers about the statistical theory of NNs, see for example the overview by Fan et al. (2021). Many of these papers attempt to mathematically explain some of the properties of NNs that have been observed in practice. An example of such a property is that NNs seem to circumvent the curse of dimensionality. Let d denote the dimension of the input. The classical minimax rate of L^q-convergence for nonparametric regression proven by Stone (1982) is $n^{-p/(2p+d)}$ for a p times differentiable regression function and $q < \infty$. This suggests that learning should be extremely slow if the input is high-dimensional, yet NNs have been found to perform exceptionally well in high-dimensional nonparametric regression problems, such as when images are used as input. Iizuka et al. (2016), for example, use grayscale images of size 224 × 224 as input for their NN, leading to 50176 regressors.

To avoid the curse of dimensionality, a different class of possible regression functions must be considered than the class of p times differentiable functions considered by Stone (1982). Barron (1994), for example, showed that shallow NNs with sigmoidal activation can achieve an L^2-convergence rate of $(n/\log n)^{-1/4}$ when estimating a regression function from the class of functions $f : [-1, 1]^d \to \mathbb{R}$ with $C_f := \int_{\mathbb{R}^d} |\omega|_1 |\tilde{F}|(d\omega) < C$, where $|\cdot|_1$ is the ℓ^1 norm, $\tilde{F}$ is the Fourier transform of f as a complex-valued measure and $C > 0$ is some known constant. This seems to avoid the curse of dimensionality, since the rate does not seem to depend on d except through a constant factor. However, as Barron (1994) already noted, C_f can indirectly depend on d because d determines the domain of integration. Furthermore, the rate $(n/\log n)^{-1/4}$ is slower than the minimax rate given by Stone (1982) if the regression function is at least $d/2$ times differentiable.

More recently, the convergence rate of deep NNs has been investigated for regression functions that can be represented as compositions of functions from certain classes. Intuitively, this might be a natural class of regression functions to consider, since deep NNs are themselves defined through compositions of layers. Bauer and Kohler (2019) follow this general idea and consider regression functions that can be represented through composition and summation of functions from certain classes. For such regression functions, they show that truncated deep NNs can achieve a $\log(n)^{3/2}n^{-p/(2p+d^*)}$ rate of convergence, if the component functions of the regression function are p-smooth and depend on at most d^* variables each. This looks strikingly similar to the rate of Stone (1982), except for the log factors and d being replaced by d^*. The latter is essential for circumventing the curse of dimensionality, since d^* is assumed to be much smaller than d in practice. However, their result requires that the activation function is bounded and at least p times differentiable. This smoothness assumption excludes some popular activation functions, such as the ReLU activation function $x \mapsto \max(0, x)$. The ReLU activation function is popular in practice due to computational reasons. Schmidt-Hieber (2020) derived similar convergence rates for deep NNs with ReLU activation and regression functions under a rather general composition assumption. In particular, the convergence rate is again determined by the smoothness of the component functions and the number of arguments they depend on.

The aim of this study is to consider the work of Schmidt-Hieber (2020) in more detail. First, we will revisit his main results and their implications. We discuss some small issues in his proofs and suggest minor modifications to his results that are sufficient to solve these issues. We will provide detailed proofs for these modified results in the appendix. We also provide a generalization of his main theorem to allow for Markov chain regressors and dependent errors. Next, we discuss some examples of applications of the results of Schmidt-Hieber (2020). Finally, we will discuss practical aspects for translating these theoretical results to an implementation in practice.

2 Theoretical framework

In this section, we first introduce the nonparametric regression model considered by Schmidt-Hieber (2020). We then discuss deep NNs and the specific constraints imposed on them by Schmidt-Hieber (2020). Next, we revisit the main results proven by Schmidt-Hieber (2020). We point out some points at which his proof is not completely clear and discuss some minor modifications to his results that are sufficient for us to complete the proof. Finally, we discuss some ways these results can be generalized. All proofs are deferred to the appendix.

All theorems and lemmas given in this thesis are due to Schmidt-Hieber (2020) and follow his numbering. All theorems and lemmas with a number not followed by a prime, such as Theorem 1, are taken directly from Schmidt-Hieber (2020) with only minor changes in wording or notation for clarity. All theorems and lemmas with a number followed by a prime, such as Theorem $1'$, are versions of the corresponding result without a prime, but with small changes to the results themselves. These changes are consequences of our solutions to certain points that are unclear to us in the proofs provided by Schmidt-Hieber (2020). The propositions are our own results, but they are clearly inspired by Schmidt-Hieber (2020).

Convention: On any topological space, we consider the Borel sigma-algebra.
Notation: Let $\mathbb{N}$ denote the positive integers and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}_0$, let $[n] := \{m \in \mathbb{N} : m \leq n\}$ and $[n]_0 := [n] \cup \{0\}$. We will denote vectors by bold symbols: $\boldsymbol{x} \in \mathbb{R}^d$. The transpose of a vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^\top$ and the p-norm of $\boldsymbol{x}$ is denoted by $|\boldsymbol{x}|_p$. For two measures μ, ν, denote their product measure by $\mu \times \nu$. For a measurable subset $D \subset \mathbb{R}^m$ and constant $p \in (0, \infty]$, $|| \cdot ||_{L^p(D)}$ denotes the L^p norm with respect to Lebesgue measure. If D is clear, we also write $|| \cdot ||_{L^p}$. For a function $f : D \to \mathbb{R}$, denote the sup-norm by $||f||_\infty := \sup_{\boldsymbol{x} \in D} |f(\boldsymbol{x})|$. For two non-negative sequences $(a_n)_n, (b_n)_n$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all n and we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For $x \geq 0$, we define $\lfloor x \rfloor := \max\{n \in \mathbb{Z} : n \leq x\}$, $\underline{x} := \max\{n \in \mathbb{Z} : n < x\}$, $\overline{x} := x - \underline{x}$ and $\lceil x \rceil := \min\{n \in \mathbb{Z} : n \geq x\}$. For functions $g : [a, b]^n \to \mathbb{R}^m$ and $g_j : [a, b]^n \to \mathbb{R}$, we mean by $g = (g_j)_{j=1}^m$ that $g(\boldsymbol{x}) = (g_j(\boldsymbol{x}))_{j=1}^m$ for all $\boldsymbol{x} \in [a, b]^n$. log denotes the natural logarithm, $\log_a$ denotes the logarithm with base a and $\log^b n = (\log n)^b$. When writing equalities or inequalities between random variables or between a random variable and a constant, we mean that they hold almost surely.

2.1 The nonparametric regression model

Schmidt-Hieber (2020) considers the following nonparametric regression model

$$Y_i = f_0(\boldsymbol{X}_i) + \epsilon_i, \quad i = 1, ..., n. \tag{1}$$

The regressors $(\boldsymbol{X}_i)_{i=1}^n$ are assumed to be iid with each $\boldsymbol{X}_i$ taking values in the unit hypercube $[0,1]^d$ and the noise variables $(\epsilon_i)_{i=1}^n$ are assumed to be iid standard normal and independent of $(\boldsymbol{X}_i)_{i=1}^n$. The sample size n is known and non-random. The regression function $f_0 : [0,1]^d \to \mathbb{R}$ is an unknown measurable function. We wish to estimate f_0 from the observed sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$.

To make this problem feasible, we must restrict the set of possible regression functions, or restrict the set of possible distributions of $\boldsymbol{X}_1$. We will take the first approach. As is common in nonparametric statistics, we will make use of the notion of Hölder smoothness.

2.1.1 Hölder smoothness

Hölder smoothness is defined as follows:

Definition 1. *Let $\beta > 0$ and $D \subset \mathbb{R}^m$. A function $f : D \to \mathbb{R}$ has Hölder smoothness (index) β if all partial derivatives of order up to and including $\underline{\beta}$ exist and the partial derivatives of order $\underline{\beta}$ are Hölder continuous with exponent $\overline{\beta}$. In that case, f is also called β-Hölder and its Hölder norm is defined by*

$$||f||_{H^\beta} := \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^m \\ |\boldsymbol{\alpha}|_1 < \beta}} ||\partial^{\boldsymbol{\alpha}} f||_\infty + \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^m \\ |\boldsymbol{\alpha}|_1 = \underline{\beta}}} \sup_{\substack{\boldsymbol{x}, \boldsymbol{y} \in D \\ \boldsymbol{x} \neq \boldsymbol{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\overline{\beta}}},$$

where we use multi-index notation for the partial derivatives. For a radius $K > 0$, the ball of functions on D with Hölder smoothness β is

$$\mathcal{C}_m^\beta(D, K) := \{f : D \to \mathbb{R} : f \text{ is } \beta\text{-Hölder and } ||f||_{H^\beta} \leq K\}.$$

Remark 2.1. *In his definition of Hölder smoothness, Schmidt-Hieber (2020) explicitly requires that the partial derivatives are bounded. We have omitted this here to be more* *in line with the related literature. This omission is irrelevant for the results of Schmidt-Hieber (2020), because we will only consider compact domains and then continuity implies boundedness.*

We will often abuse notation somewhat and write $g \in \mathcal{C}_m^\beta([a,b]^m, K)$ for a function $g : [a,b]^r \to \mathbb{R}$ with $m < r$ that only depends on m out of its r input variables. More formally, for a function $g : [a,b]^r \to \mathbb{R}$, we mean by $g \in \mathcal{C}_m^\beta([a,b]^m, K)$ that there exists a function $\tilde{g} : [a,b]^m \to \mathbb{R}$ and $v_1, ..., v_m \in [r]$ such that $\tilde{g} \in \mathcal{C}_m^\beta([a,b]^m, K)$ and $g(x_1, ..., x_r) = \tilde{g}(x_{v_1}, ..., x_{v_m})$ for all $x_1, ..., x_r \in [a,b]$.

2.1.2 Assumptions on the regression function

We are now able to define the function spaces that Schmidt-Hieber (2020) considers for the regression function:

$$\begin{aligned}\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K) :=\{&g_q \circ g_{q-1} \circ ... \circ g_1 \circ g_0 : \forall i \in [q]_0 \;\; \exists a_i, b_i \in [-K, K] \text{ such that} \\ &g_i = (g_{ij})_{j=1}^{d_{i+1}} : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}, \text{ for some} \\ &g_{ij} : [a_i, b_i]^{d_i} \to \mathbb{R} \text{ with } g_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K)\} \cap \{f : [0,1]^{d_0} \to \mathbb{R}\},\end{aligned} \tag{2}$$

where $q \in \mathbb{N}_0$, $\quad (d_i)_{i=0}^{i=q+1} := \boldsymbol{d} \in \mathbb{N}^{q+2}$, $\quad (t_i)_{i=0}^{q} := \boldsymbol{t} \in \mathbb{N}^{q+1}$, $\quad (\beta_i)_{i=0}^{q} := \boldsymbol{\beta} \in \mathbb{R}_{>0}^{q+1}$ and $K \geq 1$, with $t_i \leq d_i$ for all $i \in [q]_0$. Our definition of this space differs slightly from that of Schmidt-Hieber (2020) in that we restrict the class to functions $[0,1]^{d_0} \to \mathbb{R}$ and require $K \geq 1$, while Schmidt-Hieber (2020) does neither. We include this additional restriction because this will be the class we consider for the regression function and the regression function is assumed to be a function $f_0 : [0,1]^d \to \mathbb{R}$.

Note that for any $g_i = (g_{ij})_{j=1}^{d_{i+1}}$ as in (2), we have for all j that g_{ij} depends on at most t_i of its d_i input variables, but these t_i variables need not be the same for different $j \in [d_{i+1}]$. That is, g_i can depend on all d_i input variables, but each g_{ij} individually only depends on t_i input variables. Intuitively, t_i can be viewed as the "intrinsic input dimensionality" of the functions $(g_{ij})_{j=1}^{d_{i+1}}$, that is, of the function g_i. The idea is then that the dependence of the convergence rate on input dimension d can be replaced by dependence on $\boldsymbol{t}$ and that in practice $t_i \ll d$. This could then be a way to circumvent the curse of dimensionality. Note further that the particular case that $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ requires $d_0 = d$ and $d_{q+1} = 1$.

Unlike the conventional approach of making smoothness assumptions about f_0 directly, Schmidt-Hieber (2020) assumes that f_0 is a composition of functions and then makes smoothness assumptions about the components of f_0. Still, this indirectly implies smoothness assumptions on f_0 as well. This is summarized in Proposition 1. The proof can be found in Appendix B.

Proposition 1. *Let $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ be as in* (2) *and let $f = (f_j)_{j=1}^{d_{q+1}} \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$.*

- *If $\min_i \beta_i \leq 1$, then each f_j has Hölder smoothness index $\prod_{i \in [q]_0 : \beta_i \leq 1} \beta_i$.*
- *If $\min_i \beta_i \geq 1$, then each f_j has Hölder smoothness index $\min_i \beta_i$.*

We now discuss how the class $\mathcal{G} := \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ with $d_0 = d$ and $d_{q+1} = 1$ compares to classes for regression functions considered in related literature. For any constants $p, C > 0$, Stone (1982) considered the class of functions with Hölder smoothness index p whose $\underline{p}$-th order partial derivatives are Hölder continuous with exponent $\overline{p}$ and coefficient C. For $\boldsymbol{\beta} \in \mathbb{R}_{>0}^{q+1}$, denote by $\tilde{\beta}$ the Hölder smoothness index guaranteed by Proposition 1. Then by Theorem 1 of Stone (1982), for any $q \in (0, \infty)$ and under the appropriate regularity conditions, regression functions in $\mathcal{G}$ can be estimated with an $n^{-p/(2p+d)}$ rate of L^q-convergence. However, this does not take into account the composition structure of functions in $\mathcal{G}$, and their potentially limited "intrinsic dimensionality". This causes the rate guaranteed by Stone (1982) to suffer from the curse of dimensionality. As we shall see, Theorem 1 of Schmidt-Hieber (2020) implies that faster estimation can be possible because of this additional structure.

Bauer and Kohler (2019) also consider function classes with more structure than the very general classes of Stone (1982). In Definition 2 of Bauer and Kohler (2019), they define the (p, C)-smooth generalized hierarchical model of order d^* and level l, where $d^* \in [d], l \in \mathbb{N}_0$ and $p, C > 0$. Denote this model by $\mathcal{H}(p, C, d^*, l)$. Similar to Stone (1982), Bauer and Kohler (2019) call a function $f : \mathbb{R}^d \to \mathbb{R}$ (p, C)-smooth if it has Hölder smoothness p and the $\underline{p}$-th order partial derivatives are Hölder continuous with exponent $\overline{p}$ and coefficient C.

The (p, C)-smooth generalized hierarchical models of order d^* are defined via recursion. First, they define

$$\mathcal{H}(p, C, d^*, 0) := \{h : \mathbb{R}^d \ni \boldsymbol{x} \mapsto f(A\boldsymbol{x}) : h \text{ and } f \text{ are } (p, C)\text{-smooth, } A \in \mathbb{R}^{d^* \times d}\}.$$

Next, for $l \in \mathbb{N}_0$, the model of level $l + 1$ consists of the (p, C)-smooth functions that can be written as $\sum_{i=1}^{N}(h_i \circ f_i)$, for some $f_i = (f_{ij})_{j=1}^{d^*}$, (p, C)-smooth functions $h_i : \mathbb{R}^{d^*} \to \mathbb{R}$, and $f_{ij} \in \mathcal{H}(p, C, d^*, l)$. Theorem 1 of Bauer and Kohler (2019) imposes the additional assumption that all $\underline{p}$-th order partial derivatives of h_i and f_{ij} are bounded and that all h_i are Lipschitz. Under some additional regularity conditions, they then prove that their proposed estimator achieves a $\log^{3/2}(n)n^{-p/(2p+d^*)}$ rate of convergence. Up to logarithmic factors, this is the same rate as given by Stone (1982), but with the input dimension d replaced by the "intrinsic dimensionality" d^*. Note that $\mathcal{H}(p, C, d^*, l)$ assumes the same "intrinsic dimensionality" for each level. In $\mathcal{G}$, this would roughly correspond to choosing $t_i = d^*$ for all i. That is, the space considered by Schmidt-Hieber (2020) is more flexible in the sense of allowing each level to have a different "intrinsic dimensionality". The same applies to the smoothness, with Bauer and Kohler (2019) requiring every level to have the same smoothness p, which roughly corresponds to choosing $\beta_i = p$ for all i, whereas Schmidt-Hieber (2020) allows each level to have a different smoothness.

2.2 Neural networks

The goal of this section is to define NNs with ReLU activation. The ReLU activation function is defined as $\sigma(x) := \max(x, 0)$ for all $x \in \mathbb{R}$. For a shift vector $\boldsymbol{v} = (v_i)_{i=1}^r \in \mathbb{R}^r$, the shifted activation function $\sigma_{\boldsymbol{r}} : \mathbb{R}^r \to \mathbb{R}^r$ is defined as

$$\sigma_{\boldsymbol{v}}(\boldsymbol{x}) := \begin{pmatrix} \sigma(x_1 - v_1) \\ \dots \\ \sigma(x_r - v_r) \end{pmatrix}, \quad \text{for all } \boldsymbol{x} = (x_i)_{i=1}^r \in \mathbb{R}^r.$$

For convenience, we often omit the parentheses and write $\sigma_{\boldsymbol{v}}\boldsymbol{x} := \sigma_{\boldsymbol{v}}(\boldsymbol{x})$. For $L \in \mathbb{N}$ and $\boldsymbol{p} = (p_i)_{i=0}^{L+1} \in \mathbb{N}^{L+2}$, an NN (function) with network architecture $(L, \boldsymbol{p})$ is a function of the form

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad f(\boldsymbol{x}) = W_L \sigma_{\boldsymbol{v}_L} W_{L-1} \sigma_{\boldsymbol{v}_{L-1}} ... W_1 \sigma_{\boldsymbol{v}_1} W_0 \boldsymbol{x}, \tag{3}$$

for some weight matrices $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ and shift vectors $\boldsymbol{v}_i \in \mathbb{R}^{p_i}$. $\boldsymbol{v}_0$ is taken to be $\boldsymbol{0}$ by convention. L is called the number of hidden layers or depth and $\boldsymbol{p}$ is called the width vector, with p_i being the number of units in layer i. Note that NN functions with ReLU activation are piecewise linear, since $\sigma_{\boldsymbol{v}_i}$ is piecewise linear for all i and composition, addition and scalar multiplication of piecewise linear functions conserve piecewise linearity.

This definition for an NN function immediately makes clear the intuitive connection between NNs and the class $\mathcal{G}$ defined above. Indeed, NNs have a natural representation as a composition of functions.

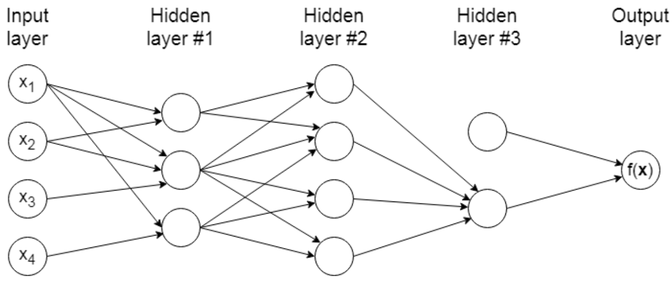

Figure 1: Example of an NN $f : \mathbb{R}^4 \to \mathbb{R}$ visualised as a layered directed acyclic graph, with $L = 3$ hidden layers and width vector $\boldsymbol{p} = (4, 3, 4, 2, 1)$.

NNs can be visualized using layered directed acyclic graphs, as shown in Figure 1. In such a visualisation, edges can only run between nodes from neighbouring layers and each vertex stands for a scalar-valued function of the input $\boldsymbol{x} \in \mathbb{R}^{p_0}$. The first layer of the network is the input layer with p_0 vertices, one for each input variable. Similarly, the final layer is the output layer with p_{L+1} vertices, one for each output variable. The L layers in between are called hidden layers and they have $p_1, p_2, ..., p_L$ vertices respectively. Each arrow stands for multiplying the tail of the arrow by a weight and feeding it as input to the head of the arrow. Each hidden layer vertex sums its input and applies a shifted ReLU activation to the sum. Each output layer vertex simply calculates the sum of its input. Thinking of NNs as graphs will often prove to be useful to us.

Let f be as in (3). Note that the arrow from the k-th vertex of layer i to the j-th vertex of layer $i+1$ has weight $(W_i)_{jk}$, and the vertex m of hidden layer l is associated with shift parameter v_{lm}. In particular, $(W_i)_{jk} = 0$ for some i, j, k means that the edge corresponding to this entry can be removed from the graph. If $(W_i)_{jk} = 0$ for some i, j and all k, then there are no arrows pointing to the j-th vertex of layer $i+1$ and this vertex simply has constant value. This corresponds to the j-th row of W_i being zero. If this is a hidden layer vertex with shift parameter 0, then this vertex has constant value 0 and can be removed from the graph.

We will consider the network architecture $(L, \boldsymbol{p})$ to be non-random hyperparameters depending on n and we will consider the network parameters $W_0, ..., W_L$ and $\boldsymbol{v}_1, ..., \boldsymbol{v}_L$ to be parameters to be chosen through some data-based estimation method. In practice, it is often observed that the network parameters in trained ReLU NNs are small. Therefore, Schmidt-Hieber (2020) chooses to bound the network parameters in absolute value by one. The space of network functions with network architecture $(L, \boldsymbol{p})$ and network parameters bounded in absolute value by one is then

$$\mathcal{F}(L, \boldsymbol{p}) := \{f : f \text{ is of the form (3) with } ||W_j||_\infty, |\boldsymbol{v}_j| \leq 1 \text{ for all } j \in [L]_0\},$$

where $||W||_\infty := \max_{i \in [k_1], j \in [k_2]} |W_{ij}|$ for any matrix $W \in \mathbb{R}^{k_1 \times k_2}$.

As a form of regularization, Schmidt-Hieber (2020) bounds the number of non-zero parameters by some $s \in \mathbb{N}_0$. s is again a non-random hyperparameter that depends on n. The idea is to prevent overfitting by choosing s small compared to the total number of network parameters. Let $||W||_0$ be the number of non-zero entries of a matrix W. Then in $\mathcal{F}(L, \boldsymbol{p})$, the s-sparse NNs bounded on $[0,1]^{p_0}$ by $F > 0$ are

$$\mathcal{F}(L, \boldsymbol{p}, s, F) := \{f \in \mathcal{F}(L, \boldsymbol{p}) : \sum_{j=0}^{L} ||W_j||_0 + |\boldsymbol{v}_j|_0 \leq s, \sup_{\boldsymbol{x} \in [0,1]^{p_0}} |f(\boldsymbol{x})|_\infty \leq F\}.$$

We further define spaces of NNs restricted to $[0,1]^{p_0}$:

$$\mathcal{F}_0(L,\boldsymbol{p}) := \{f|_{[0,1]^{p_0}} : f \in \mathcal{F}(L,\boldsymbol{p})\}, \quad \mathcal{F}_0(L,\boldsymbol{p},s,F) := \{f|_{[0,1]^{p_0}} : f \in \mathcal{F}(L,\boldsymbol{p},s,F)\}.$$

We will call the functions in these spaces NNs as well.

To estimate the regression function f_0 from (1), we will consider estimators $\hat{f}_n$ that take value in $\mathcal{F}_0(L,\boldsymbol{p},s,F)$ with $p_0 = d$, $p_{L+1} = 1$. Although we consider a nonparametric regression model, we note that for a fixed choice of $L,\boldsymbol{p},s$, the estimator $\hat{f}_n$ can be seen as a parametric estimator. However, the idea will be to fix F and to let $L,\boldsymbol{p}$ and s grow at some appropriate rate as $n \to \infty$. That is, we study the estimation of f_0 through sieve estimation with NNs.

We will focus on estimators $\hat{f}_n$ that attempt to minimize the mean squared residual $n^{-1}\sum_{i=1}^n (Y_i - \hat{f}_n(\boldsymbol{X}_i))^2$. This choice of empirical risk is popular in practice and a natural choice under the assumption that the errors are normal. However, actual empirical risk minimization is not always feasible in practice, nor always desired. Hence, we consider

$$\Delta_n(\hat{f}_n, f_0) := \mathbb{E}_{f_0}\left[\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{f}_n(\boldsymbol{X}_i))^2 - \inf_{f\in\mathcal{F}_0(L,\boldsymbol{p},s,F)} \frac{1}{n}\sum_{i=1}^n (Y_i - f(\boldsymbol{X}_i))^2\right] \tag{4}$$

as a measure of how far $\hat{f}_n$ is in expectation from being an actual empirical risk minimizer. In particular, $\Delta_n(\hat{f}_n, f_0) \geq 0$ always and $\Delta_n(\hat{f}_n, f_0) = 0$ if and only if $\hat{f}_n$ is an empirical risk minimizer almost surely. To evaluate the estimator $\hat{f}_n$, we consider the following risk function:

$$R(\hat{f}_n, f_0) := \mathbb{E}_{f_0}\left[(\hat{f}_n(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right],$$

where $\boldsymbol{X}$ has the same distribution as $\boldsymbol{X}_1$ and is independent of the sample $(\boldsymbol{X}_i, Y_i)_i$. That is, we consider the expected squared prediction error. Note that this can also be seen as (expected) L^2-distance $\mathbb{E}_{f_0}\left[\int_{[0,1]^d} (\hat{f}_n(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2 d\mu(\boldsymbol{x})\right]$, where μ is the distribution of $\boldsymbol{X}_1$.

2.3 Rate of convergence for sparse neural networks

Theorem 1 of Schmidt-Hieber (2020) describes an upper bound on the expected squared prediction error for an estimator $\hat{f}_n$ taking values in $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$. Let $q, \boldsymbol{\beta}, \boldsymbol{t}$ be as in (2). To state the theorem, we must first define

$$\beta_i^* := \beta_i \prod_{j=i+1}^{q} (\beta_j \wedge 1) \quad \text{for all } i \in [q]_0 \quad \text{and} \quad \phi_n := \max_{i \in [q]_0} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}}.$$

This now allows us to state Theorem 1 of Schmidt-Hieber (2020). Although we omit this from notation, as Schmidt-Hieber (2020) did, the parameters $L, \boldsymbol{p}$ and s in the statement of Theorem 1 should be thought of as depending on n.

Theorem 1. *Consider the d-variate nonparametric regression model* (1). *Suppose* $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ *for some* $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ *as in* (2). *Let* $C_{(ii)}, C_{(iii)}, C_{(iv)}^-, C_{(iv)}^+ > 0$ *be constants and let* $L > 0, \boldsymbol{p} = (p_i)_{i=0}^{L+1} \in \mathbb{N}^{q+2}$ *and* $s \in \mathbb{N}$, *with* $p_0 = d, p_{L+1} = 1$. *Let* $F \geq 1$ *and let* $\hat{f}_n$ *be an estimator taking values in* $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$. *Suppose the following hold:*

(i) $F \geq K$,

(ii) $\sum_{i=0}^{q} \log_2(4t_i \vee 4\beta_i) \log_2(n) \leq L \leq C_{(ii)} n\phi_n$,

(iii) $n\phi_n \leq C_{(iii)} \min_{i \in [L]} p_i$,

(iv) $C_{(iv)}^- n\phi_n \log n \leq s \leq C_{(iv)}^+ n\phi_n \log n$.

Then there exist constants $C, C' > 0$ *depending only on* $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F$ *such that if* $\Delta_n(\hat{f}_n, f_0) \leq C\phi_n L \log^2 n$, *then*

$$R(\hat{f}_n, f_0) \leq C' \phi_n L \log^2 n \tag{5}$$

and if $\Delta_n(\hat{f}_n, f_0) \geq C\phi_n L \log^2 n$, *then*

$$\frac{1}{C'} \Delta_n(\hat{f}_n, f_0) \leq R(\hat{f}_n, f_0) \leq C' \Delta_n(\hat{f}_n, f_0). \tag{6}$$

Remark 2.2. *Schmidt-Hieber (2020) states his Theorem 1 without introducing the constants* $C_{(ii)}, C_{(iii)}, C_{(iv)}^-, C_{(iv)}^+$. *Instead, Schmidt-Hieber (2020) states (ii)-(iv) using the symbols* $\lesssim$ *and* $\asymp$. *This is clearly equivalent, but the constants* $C_{(ii)}, C_{(iii)}, C_{(iv)}^-, C_{(iv)}^+$ *make our discussion of Theorem 1 easier.*

Two points in the proof of Theorem 1 provided by Schmidt-Hieber (2020) are unclear to us. First of all, Schmidt-Hieber (2020) seems to claim that the constants C, C' can be chosen independent of $C_{(ii)}, C_{(iii)}, C^{-}_{(iv)}, C^{+}_{(iv)}$. However, already in his (22) at the start of his proof of Theorem 1, it seems that he uses that $s \leq C'' n\phi_n \log n$, for some constant $C'' > 0$ depending only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F$. We do not see why this is guaranteed by (iv). He seems to use (ii) and (iii) of Theorem 1 in a similar way. To solve this problem in the proofs, we therefore suspect that the constants C, C' should also be allowed to depend on $C_{(ii)}, C_{(iii)}, C^{-}_{(iv)}, C^{+}_{(iv)}$.

In fact, Theorem 1 is not true without this modification. Indeed, suppose that $C' > 0$ is the constant from Theorem 1, depending only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F$, and suppose $f_0 \equiv c \neq 0$ is some non-zero constant function. Then for n large enough, $C'\phi_n \log^2 n < c^2$. However, by choosing appropriate constants $C_{(ii)}, C_{(iii)}, C^{-}_{(iv)}, C^{+}_{(iv)}$, the network hyperparameters $L = 1, \boldsymbol{p} = (d, 1, 1), s = 1$ satisfy (i)-(iv) of Theorem 1 for this n. In that case, $\mathcal{F}_0(L, \boldsymbol{p}, s)$ consists only of the zero function and hence $R(\hat{f}_n, f_0) = \mathbb{E}[f_0(\boldsymbol{X})^2] = c^2$. This is a contradiction with (5) and the fact that $C'\phi_n \log^2 n < c^2$.

Secondly, to obtain his (26), Schmidt-Hieber (2020) seems to implicitly use that $Nn^{-1} \lesssim N^{-\beta_i/t_i}$ for some $i \in [q]_0$, where $N = \lceil c \max_{j\in[q]_0} n^{t_j/(2\beta_j^*+t_j)} \rceil$ for some sufficiently small $c > 0$. We do not see how this follows from the assumptions in Theorem 1. Indeed, for any $\ell \in \operatorname{argmax}_{j\in[q]_0} n^{t_j/(2\beta_j^*+t_j)}$ and any $i \in [q]_0$, we have

$$Nn^{-1} \asymp n^{-\frac{2\beta_\ell^*}{2\beta_\ell^*+t_\ell}}, \quad N^{-\beta_i/t_i} \asymp n^{-\frac{-\beta_i}{2\beta_\ell^*+t_\ell}\cdot\frac{t_\ell}{t_i}}.$$

It is then easy to construct counterexamples for which we do not have $Nn^{-1} \lesssim N^{-\beta_i/t_i}$ for any $i \in [q]_0$. Consider for example $q = 3$ with $\beta_0 = 1, \beta_1 = \beta_2 = \beta_3 = 1/4$ and $t_j = 1$ for all $j \in [q]_0$. To solve this problem, we slightly modify the proof and replace (ii) in Theorem 1.

Combining the above two points yields Theorem $1'$.

Theorem 1′. *Theorem 1 holds if the constants C, C' are also allowed to depend on*

$C_{(ii)}, C_{(iii)}, C^-_{(iv)}, C^+_{(iv)}$, and we replace (ii) by

(ii') $\sum_{i=0}^q \frac{\beta_i+t_i}{2\beta_i^*+t_i}\log_2(4t_i \vee 4\beta_i)\log_2(n) \leq L \leq C_{(ii)} n\phi_n$.

Note that (ii') can be a weaker or stronger condition on L than (ii), depending on the vectors $\boldsymbol{\beta}$ and $\boldsymbol{t}$.

Fortunately, our interpretation of Theorem $1'$ is the same as that of Theorem 1, since the changes are relatively minor. Under the appropriate assumptions, the theorem essentially states that if the estimator is too far from being an empirical risk minimizer, then $\Delta_n(\hat{f}_n, f_0)$ determines the rate of convergence: $R(\hat{f}_n, f_0) \asymp \Delta_n(\hat{f}_n, f_0)$. However, if $\hat{f}_n$ is sufficiently close to being an empirical risk minimizer, then the rate of convergence is determined by ϕ_n and L: $R(\hat{f}_n, f_0) \leq C'\phi_n L \log^2 n$. In particular, Schmidt-Hieber (2020) gives the following direct consequence of Theorem 1:

Corollary 1. *Suppose the same conditions hold as in Theorem 1. Let $\tilde{f}_n$ be an estimator taking values in $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$ and suppose $\tilde{f}_n$ is an empirical risk minimizer. That is, $\tilde{f}_n \in \operatorname{argmin}_{f\in\mathcal{F}_0(L,\boldsymbol{p},s,F)} \sum_{i=1}^n (Y_i - f(\boldsymbol{X}_i))^2$ almost surely. Then*

$$R(\tilde{f}_n, f_0) \leq C'\phi_n L \log^2 n,$$

where $C' > 0$ is the same constant as in the statement of Theorem 1.

Of course if Theorem $1'$ is assumed to be true instead of Theorem 1, then Corollary 1 still follows, but with Theorem 1 replaced by Theorem $1'$. The detailed proof of Theorem $1'$ can be found in Appendix A. The idea is to combine results on approximation of functions by ReLU NNs and the following theorem of Schmidt-Hieber (2020):

Theorem 2. *Consider the d-variate nonparametric regression model (1) with $||f_0||_\infty \leq F$ for some $F \geq 1$. Let $L > 0, \boldsymbol{p} = (p_i)_{i=0}^{L+1} \in \mathbb{N}^{q+2}$ and $s \geq 1$, with $p_0 = d, p_{L+1} = 1$. Let $\hat{f}_n$ be an estimator taking values in $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$. For any $\epsilon \in (0, 1]$, there exists a constant C_ϵ depending only on ϵ such that*

$$\begin{aligned}(1-\epsilon)^2\Delta_n(\hat{f}_n, f_0) - \tau_{\epsilon,n} &\leq R(\hat{f}_n, f_0) \\ &\leq (1+\epsilon)^2\left(\inf_{f\in\mathcal{F}_0(L,\boldsymbol{p},s,F)} ||f-f_0||_\infty^2 + \Delta_n(\hat{f}_n, f_0)\right) + \tau_{\epsilon,n},\end{aligned}$$

where

$$\tau_{\epsilon,n} := C_\epsilon F^2 n^{-1}(s+1)\log(n(s+1)^L d).$$

The proof Schmidt-Hieber (2020) gives of Theorem 2 is based on his Lemma 4, which can also be found in our Appendix A. However, his proof of Lemma 4 contains an application of Bernstein's inequality that seems incorrect to us. More details are given in Remark A.2 in Appendix A. To solve this problem, we again modify the proof slightly, leading to a slightly different version of Lemma 4, which we call Lemma $4'$. Lemma $4'$ then leads to a slightly different version of Theorem 2, which we call Theorem $2'$.

Theorem $2'$. *Theorem 2 holds if we replace the definition of $\tau_{\epsilon,n}$ by*

$$\tau_{\epsilon,n} := C_\epsilon F^4 n^{-1}(s+1)\log(n(s+1)^L d).$$

Again, the difference between Theorems 2 and $2'$ is only minor and it does not affect the proof of Theorem $1'$. The proofs of Lemma $4'$ and Theorem $2'$ can be found in Appendix A.

The approximation theory for ReLU networks is based on two key properties of the ReLU activation function σ. First of all, σ is a projection: $\sigma \circ \sigma = \sigma$. This means that input can pass through several layers unchanged. In particular, if we have input dimension d and layer widths $p_i = d$ for all i, then choosing shift vectors 0 and identity weight matrices results in the identity as a network function. This seems simple, but for other activation functions this can be much harder. Furthermore, ReLU activation allows function approximation with network parameters bounded in absolute value by one. As an example, consider the approximation of monomials, which is essential for approximating functions through a Taylor approximation. Schmidt-Hieber (2020) shows in Lemma A.4 in his supplement that ReLU networks can approximate monomials to arbitrary accuracy with network parameters bounded in absolute value by one. On the other hand, Bauer and Kohler (2019) consider approximation of monomials by NNs with sigmoidal activation in their Lemma 4. Inspection of the proof reveals that letting the upper bound on the approximation error vanish requires letting some network parameters diverge to infinity. In their notation: the bound on the approximation error scales as R^{-1}, while the network parameter γ_k scales as R^N, where N is an upper bound on the degree of the monomial. However, this example does of course not prove that it is impossible to approximate functions to arbitrary accuracy with an NN with sigmoidal activation and a bound on the network parameters. This could be an interesting subject for future research.

Condition (i) of Theorem 1 implies that the supremum norm bound on the network functions must not be smaller than the bound on the Hölder norm of the components of the regression function. Since the Hölder norm is an upper bound on the supremum norm, this is a natural condition to require. Condition (ii) dictates that the order of L must be between $\log n$ and $n\phi_n$. Note that if $i^* \in \operatorname{argmax}_{i\in[q]_0} n^{-2\beta^*_{i^*}/(2\beta^*_{i^*}+t_{i^*})}$, then $n\phi_n = n^{t_{i^*}/(2\beta^*_{i^*}+t_{i^*})}$ and hence $n\phi_n/\log n \to \infty$. Condition (iii) states that the minimal hidden layer width must be at least of order $n\phi_n$. In particular, conditions (ii) and (iii) together mean that the network should become both deeper and wider as n increases. Note that condition (ii) imposes a lower and upper bound on the order of L, while condition (iii) only imposes a lower bound on the minimal hidden layer width. Hence, it suffices to choose the hidden layer widths sufficiently large, for example $\min_{i\in[L]} p_i \asymp n$. Condition (iv) requires that s increases with a rate $n\phi_n \log n$. Note that for a given network architecture $(L, \boldsymbol{p})$, there are $\sum_{i=0}^{L} p_i p_{i+1}$ weight parameters, which is lower bounded by $(L-1)\left(\min_{i\in[L]} p_i\right)^2$. Hence by conditions (ii) and (iii), the total number of network parameters is at least of order $n^2\phi_n^2 \log n$. Since $n\phi_n \log n/(n^2\phi_n^2 \log n) \to 0$, this implies that the fraction of the network parameters that is non-zero goes to 0. That is, the network becomes relatively more sparse as n increases. Furthermore, the rate $s \asymp n\phi_n \log n$ is at least a polynomial factor slower than n, so that also $s/n \to 0$. This means that although the class $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$ will eventually have much more parameters than n, the number of non-zero parameters will eventually be much smaller than n. As noted before, Schmidt-Hieber (2020) introduced this sparsity constraint as a type of regularization to prevent overfitting.

We now consider the implications of Theorem 1 and its corollary in more detail. First of all, the upper bound on $R(\hat{f}_n, f_0)$ is asymptotically minimized, up to a constant factor, by choosing $L \asymp \log n$, leading in particular to $R(\hat{f}_n, f_0) \lesssim \phi_n \log^3 n$. This rate does not depend directly on d. Instead, ϕ_n depends only on $\boldsymbol{\beta}$ and $\boldsymbol{t}$. One consequence of this observation is that including additional irrelevant regressors changes the upper bound $C'\phi_n L \log^2 n$ at most by a multiplicative constant, because only $d = d_0$ changes and not $\boldsymbol{t}$ or $\boldsymbol{\beta}$. This is one example of how NNs may be able to avoid the curse of dimensionality in nonparametric regression. More examples will be discussed in Section 3.

In the special case that there exists $d^* \in \mathbb{N}$ and $\beta > 0$ such that $t_i = d^*$ for all i and $\beta_i = \beta$ for all i, we have that $\phi_n = n^{-2\beta^*/(2\beta^*+d^*)}$, where $\beta^* = \beta$ if $\beta \geq 1$, and $\beta^* = \beta^{q+1}$ otherwise. Recall that in that case, the class $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ roughly corresponds to the (β, K)-smooth generalized hierarchical interaction model of order d^* and finite level q of Bauer and Kohler (2019). Under some regularity conditions, Theorem 1 of Bauer and Kohler (2019) guarantees that regression functions in this model can be estimated with a $n^{-2\beta/(2\beta+d^*)} \log^3 n$ rate for the expected squared prediction error. If $\beta \geq 1$, this is the same as the rate $\phi_n \log^3 n$ guaranteed by Theorem 1 of Schmidt-Hieber (2020). However, if $\beta < 1$, the rate guaranteed by Bauer and Kohler (2019) is at least a polynomial factor faster than $\phi_n \log^3 n$. Hence, Theorem 1 of Schmidt-Hieber (2020) seems to be of interest especially for the case that not all β_i are equal, or not all t_i are equal.

Note that Theorem 1 gives an upper bound on the expected squared prediction error, but no lower bound in the case that $\Delta_n(\hat{f}_n, f_0) \leq C\phi_n L \log^2 n$. Hence, Theorem 1 raises the question whether a better rate of convergence is possible, and if so, how much improvement is possible. Theorem 3 of Schmidt-Hieber (2020) is a partial answer to this question in the case that $t_i \leq \min_{j \in [i-1]_0} d_j$ for all i.

Theorem 3. *Consider the nonparametric regression model* (1)*, with* $\boldsymbol{X}_1$ *drawn from a distribution with a Lebesgue density on* $[0,1]^d$*, which is lower and upper bounded by positive constants. Let* $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}$ *be as in* (2)*. Let* $K \geq 1$ *be a sufficiently large constant, let* $d_0 = d, d_{q+1} = 1$ *and suppose* $t_i \leq \min_{j \in [i-1]_0} d_j$ *for all* $i \in [q]_0$*. Then there exists a positive constant* c *such that for all* n*,*

$$\inf_{\hat{f}_n} \sup_{f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)} R(\hat{f}_n, f_0) \geq c\phi_n,$$

where $\inf_{\hat{f}_n}$ *is over all estimators* $\hat{f}_n$*.*

Hence, in this particular context, the rate of Theorem 1 is optimal up to log factors. However, this is a bound on the minimax rate for the whole class $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ and an estimator could achieve a better rate for a particular regression function or a particular subset of $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$.

As Schmidt-Hieber (2020) notes, Theorems 2 and 3 can be combined to obtain lower bounds for how many non-zero parameters are required to be able to approximate a function. Lemma 1 of Schmidt-Hieber (2020) is an example of such a result.

Lemma 1. *Let $\beta, K > 0, d \in \mathbb{N}$. Then there exist constants $c_1, c_2 > 0$ depending only on β, K and d such that if*

$$s \leq c_1 \frac{\epsilon^{-d/\beta}}{L \log(1/\epsilon)},$$

for some $\epsilon \in (0, c_2]$ and $L \in \mathbb{N}$, then for any width vector $\boldsymbol{p} = (p_i)_{i=0}^{L+1}$ with $p_0 = d$ and $p_{L+1} = 1$, we have

$$\sup_{f_0 \in \mathcal{C}_d^\beta([0,1]^d, K)} \inf_{f \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty)} ||f - f_0||_\infty \geq \epsilon. \tag{7}$$

Remark 2.3. *Schmidt-Hieber (2020) does not place any restrictions on K in Lemma 1, other than positivity. However, the proof relies on applying Theorem 3, which requires K to be sufficiently large. We suspect that the omission of this requirement is an error in the statement of the lemma.*

The detailed proof of Lemma 1 can be found in Appendix A. The idea is that if it were possible to obtain a better approximation error with the same amount of non-zero parameters, then Theorem 2 would guarantee a faster rate of convergence then Theorem 3 allows. The proof given by Schmidt-Hieber (2020) for Lemma 1 is not affected by replacing Theorem 2 by Theorem 2′.

The same strategy can be used to obtain lower bounds for the L^2-approximation error and leads to Proposition 2.

Proposition 2. *Let $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ be as in (2), with $d_{q+1} = 1$, K large enough and $t_i \leq \min_{j \in [i-1]_0} d_j$ for all $i \in [q]_0$. Let $F > 0$ be a constant with $F \geq K$. Let μ be a measure on $[0,1]^{d_0}$ with a Lebesgue density bounded from above and below by positive constants. Let*

$i^* \in \operatorname{argmax}_{i\in[q]_0} n^{-2\beta_i^*/(2\beta_i^*+t_i)}$. *Then there exist constants* $c_1, c_2 > 0$ *depending only on* $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F$ *and* μ *such that if*

$$s \leq c_1 \frac{\epsilon^{-t_{i^*}/\beta_{i^*}^*}}{\log\left(\epsilon^{-1} L \prod_{\ell=0}^{L+1}(p_\ell + 1)\right)},$$

for some $\epsilon \in (0, c_2], L \in \mathbb{N}$ *and width vector* $\boldsymbol{p} = (p_i)_{i=0}^{L+1}$ *with* $p_0 = d_0$ *and* $p_{L+1} = 1$, *then we have*

$$\sup_{f_0 \in \mathcal{G}(q,\boldsymbol{d},\boldsymbol{t},\boldsymbol{\beta},K)} \inf_{f\in\mathcal{F}_0(L,\boldsymbol{p},s,F)} ||f - f_0||_{L^2([0,1]^{d_0},\mu)} \geq \epsilon.$$

The proof can be found in Appendix B and relies on combining Theorem 3 and Lemma 4′. Replacing Lemma 4′ by Lemma 4 of Schmidt-Hieber (2020) in the proof of Proposition 2 does not change the result. Note that if μ is the distribution of $\boldsymbol{X}$, the L^2-distance in Proposition 2 can also be written as $\mathbb{E}\left[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right]$.

One striking difference between Proposition 2 and Lemma 1 is that the former concerns approximation of functions in $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$, whereas the latter only concerns functions in $\mathcal{C}_d^\beta([0,1]^d, K)$. However, a closer inspection of the proof of Lemma 1 reveals that little to no modification of the proof is required to replace the space $\mathcal{C}_d^\beta([0,1]^d, K)$ by $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$, and replace $\epsilon^{-d/\beta}$ by $\epsilon^{-t_{i^*}/\beta_{i^*}^*}$, as in Proposition 2. Other than c_1, c_2 being allowed to depend on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$, no further modifications to the statement of Lemma 1 are necessary.

2.4 Generalizations

Theorems 1 and 2 of Schmidt-Hieber (2020) assume that the errors in (1) are iid standard normal, that the regressors are also iid and that the errors and regressors are independent of each other. These are rather strict assumptions that might often be unreasonable in practice. This section describes ways to relax these assumptions. We will show that Theorem 1′ also holds if $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ is a sufficiently nice Markov chain and $(\epsilon_i)_{i=1}^n$ are jointly conditionally ν-subgaussian given $(\boldsymbol{X}_i)_{i=1}^n$, for some variance proxy $\nu^2 > 0$.

In the rest of this section, we will first discuss some theoretical preliminaries on Markov chains and subgaussianity. We will then state the generalization of Theorem 1′ and discuss this result. All proofs are deferred to Appendix B.

2.4.1 Markov chains on Polish spaces

A closer inspection of the proofs of Theorems $1'$ and $2'$ reveals that these proofs rely on the assumption of iid regressors only through Lemma $4'$. The proof of Lemma $4'$, in turn, mainly relies on the iid assumption to be able to apply Bernstein's inequality to bounded functions of $(\boldsymbol{X}_i, \boldsymbol{X}_i')$, where $(\boldsymbol{X}_i')_{i=1}^n$ is a sequence of random vectors independent of and with the same distribution as $(\boldsymbol{X}_i)_{i=1}^n$. It turns out that Bernstein's inequality in this setting holds with different constants if $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ is a Markov chain with appropriate assumptions. The key to this result is Theorem 3.4 of Paulin (2015). To make it clear what the appropriate assumptions are, we will now give a brief review of time-homogeneous Markov chains on general Polish state spaces. This review will be based on Paulin (2015), Meyn and Tweedie (2012) and Douc et al. (2018).

Remark 2.4. *As Fan et al. (2021) point out in Footnote 3, the proofs in Paulin (2015) were not completely correct. Paulin has corrected this in a new version of his paper available on arXiv. Fortunately, Theorem 3.4 of Paulin (2015) remains true unchanged. Note that the numbering of results in the new version is different from the numbering in the original paper.*

We begin by defining probability kernels and Markov transition kernels.

Definition 2. *A probability kernel between two measurable spaces $(T, \mathcal{T})$ and $(S, \mathcal{S})$ is a function $P : T \times \mathcal{S} \to [0, 1]$ that satisfies the following two properties:*

- *For all $t \in T$, $P(t, \cdot)$ is a probability measure on $(S, \mathcal{S})$.*
- *For all $A \in \mathcal{S}$, $P(\cdot, A)$ is a measurable function on $(T, \mathcal{T})$.*

A (Markov) transition kernel on a Polish space $\mathcal{X}$ is a probability kernel between $(\mathcal{X}, \mathcal{B})$ and $(\mathcal{X}, \mathcal{B})$, where $\mathcal{B}$ is the Borel sigma-algebra of $\mathcal{X}$. $\mathcal{X}$ is called the state space of P.

If P and Q are transition kernels on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, then denote by $P \times Q$ the transition kernel on $\mathcal{X} \times \mathcal{Y}$ defined by $(P \times Q)((z, z'), \cdot) := P(z, \cdot) \times Q(z', \cdot)$ for all $z \in \mathcal{X}, z' \in \mathcal{Y}$.

The terms probability kernel and (Markov) transition kernel are often used synonymously in the literature to refer to either one of the concepts introduced in the above definition. We distinguish them here for notational convenience.

We can now introduce Markov chains and some basic related concepts.

Definition 3. *Let $(\mathcal{X}, \mathcal{B})$ be a Polish space with its Borel sigma-algebra. Let μ be a probability distribution on $\mathcal{X}$ and P a transition kernel on $\mathcal{X}$. A sequence $(\Phi_i)_{i=1}^{\infty}$ of random variables taking values in $\mathcal{X}$ is a (time-homogeneous) Markov chain with transition kernel P and initial distribution μ if Φ_1 has distribution μ and for any $m \in \mathbb{N}$ and $A_1, ..., A_m \in \mathcal{B}$, the finite-dimensional distribution is given by*

$$\begin{aligned}&\mathbb{P}\left(\Phi_1 \in A_1, \Phi_2 \in A_2, ..., \Phi_m \in A_m\right)\\ &= \int_{A_1}\int_{A_2} ... \int_{A_{m-1}} P(z_{m-1}, A_m) P(z_{m-2}, dz_{m-1}) ... P(z_1, dz_2)\mu(dz_1).\end{aligned}$$

The m-step transition kernel for a transition kernel P on $\mathcal{X}$ and $m \in \mathbb{N}$ is the transition kernel on $\mathcal{X}$ defined recursively by

$$P^0(z, A) := \delta_z(A), \quad P^m(z_1, A) := \int_{\mathcal{X}} P(z_m, A) P^{m-1}(z_1, dz_m),$$

where δ_z denotes the Dirac measure at z.

A probability distribution π on $(\mathcal{X}, \mathcal{B})$ is a stationary distribution for the transition kernel P if

$$\pi(A) = \int_{\mathcal{X}} P(z, A)\pi(dz), \quad \textit{for all } A \in \mathcal{B}.$$

A Markov chain with transition kernel P and unique stationary distribution π is called periodic if there exists an integer $m \geq 2$ and disjoint subsets $U_1, ..., U_m \in \mathcal{B}$ with $\pi(U_i) > 0$ for all i, such that $P(z, U_{i+1}) = 1$ for all $i \in [m-1]$ and $z \in U_i$, and such that $P(z, U_1) = 1$ for all $z \in U_m$. Otherwise, P is called aperiodic.

A Markov chain with transition kernel P is called ϕ-irreducible if there exists a non-zero, σ-finite measure λ on $\mathcal{X}$ such that for all $A \in \mathcal{B}$ with $\lambda(A) > 0$ and for all $z \in \mathcal{X}$, there exists $m \in \mathbb{N}$ such that $P^m(z, A) > 0$. P is then called ϕ-irreducible with $\phi = \lambda$.

A transition kernel P with stationary distribution π is uniformly ergodic if there exist $\rho \in (0, 1)$ and $M > 0$ such that

$$\sup_{z \in \mathcal{X}} d_{TV}(P^m(z, \cdot), \pi) \leq M\rho^m, \quad \forall m \in \mathbb{N},$$

where $d_{TV}(P^m(z, \cdot), \pi) := \sup_{A \in \mathcal{B}} |P^m(z, A) - \pi(A)|$ denotes the total variation distance.

A sequence $(\Phi_i)_{i=1}^{\infty}$ of random variables taking values in $\mathcal{X}$ is strictly stationary if for each n, we have that the joint distributions of $(\Phi_i)_{i=1}^{n}$ and $(\Phi_i)_{i=k}^{k+n-1}$ are equal for all $k \in \mathbb{N}$.

Remark 2.5. *All of the definitions in Definition 3 are standard in the literature, except for aperiodicity. Our definition of aperiodicity is the one given by Paulin (2015) with the additional condition that the stationary distribution is unique. Paulin (2015) does not require this unicity, but his definition therefore has the drawback of being dependent on the choice of stationary distribution π. Meyn and Tweedie (2012) and Douc et al. (2018) give a more involved definition of aperiodicity that does not require the existence of a stationary distribution, but they require that the chain is ϕ-irreducible. We follow the definition of Paulin (2015) because we wish to apply his Theorem 3.4.*

Note that a Markov chain is strictly stationary if and only if its initial distribution is a stationary distribution for its transition kernel, see for example Theorem 1.4.2 of Douc et al. (2018). Furthermore, it is well-known that a ϕ-irreducible transition kernel has at most one stationary distribution, see for example Corollary 9.2.16 of Douc et al. (2018).

Paulin (2015) assumes in his entire Section 3 that the transition kernel under consideration is ϕ-irreducible and aperiodic with some unique stationary distribution π. For notational convenience we call such a kernel $\mathrm{N}(\pi)$, where the N stands for nice.

In the case that $\mathcal{X}$ is finite with k elements, a function $f : \mathcal{X} \to \mathbb{C}$ can be seen as a vector in $\mathbb{C}^k$, and a transition kernel P on $\mathcal{X}$ can be seen as a $(k \times k)$ matrix of transition probabilities. P can then be seen as a linear operator on the space $\mathbb{C}^k$ of functions $f : \mathcal{X} \to \mathbb{C}$, with $Pf \in \mathbb{C}^k$ defined by matrix-vector multiplication. To obtain results like Theorem 3.4 of Paulin (2015), we generalize this idea to the setting where $\mathcal{X}$ is a Polish space rather than a finite set. We will view a transition kernel P with stationary distribution π as a linear operator on the Hilbert space $L^2(\pi, \mathbb{C})$. This linear operator is defined as follows.

Definition 4. *For a transition kernel P on a Polish space $\mathcal{X}$ with stationary distribution π, define a linear operator $\boldsymbol{P}$ on the Hilbert space $L^2(\pi) := L^2(\pi, \mathbb{C})$ by*

$$\boldsymbol{P} : L^2(\pi) \to L^2(\pi), \quad \boldsymbol{P}[f]_\pi := \left[z \mapsto \int_{\mathcal{X}} f(x) P(z, dx) \right]_\pi ,$$

where by $[f]_\pi$ we mean the equivalence class of measurable functions $g : \mathcal{X} \to \mathbb{C}$ with $f = g$ π-a.e.

Although the definition of $\boldsymbol{P}$ depends on the choice of stationary distribution π, the particular choice of stationary distribution will always be clear from context and we omit this dependence from notation. Recall in particular that for ϕ-irreducible transition kernels, the invariant distribution is unique and there can therefore be no ambiguity at all.

It it straightforward to confirm that $\boldsymbol{P}$ is a well-defined linear operator on $L^2(\pi)$ with operator norm 1. Furthermore, it is well-known that the Hilbert adjoint $\boldsymbol{P}^*$ of $\boldsymbol{P}$ corresponds to a transition kernel P^* that also has stationary distribution π, see for example (3.2) of Paulin (2015). Similarly, if Q is also a transition kernel with stationary distribution π, then the linear operator $\boldsymbol{PQ}$ corresponds to a transition kernel PQ with stationary distribution π. PQ is given by $PQ(x, A) = \int_{\mathcal{X}} P(y, A) Q(x, dy)$. That is, the chain first transitions according to Q and then according to P. This also implies for any $m \in \mathbb{N}$ that the linear operator corresponding to P^m is the m-fold composition of $\boldsymbol{P}$ with itself.

It turns out that the spectral analysis of the operator $\boldsymbol{P}$ is intimately related to the ergodic properties of the Markov chain. The following definition defines two important quantities for spectral analysis of Markov chains: the spectral gap and pseudo spectral gap.

Definition 5. *Suppose P is a transition kernel on a Polish space $\mathcal{X}$ with stationary distribution π. P is reversible if $\boldsymbol{P}$ is self-adjoint. For a Hilbert space $\mathcal{H}$ and linear operator $T : \mathcal{H} \to \mathcal{H}$, let $\operatorname{Spec}(T) := \{\lambda \in \mathbb{C} : (T - \lambda)$ is not boundedly invertible$\}$ denote the spectrum of T. If P is reversible, the spectral gap is defined as*

$$\gamma(\boldsymbol{P}) := \begin{cases} 1 - \sup\left(\operatorname{Spec}(\boldsymbol{P}) \setminus \{1\}\right) & \textit{if the kernel of } \boldsymbol{P} - 1 \textit{ has dimension 1.} \\ 0 & \textit{otherwise.} \end{cases}$$

For not-necessarily reversible P, the pseudo spectral gap is defined as

$$\gamma_{ps}(\boldsymbol{P}) := \sup_{k \in \mathbb{N}} \frac{\gamma((\boldsymbol{P}^*)^k \boldsymbol{P}^k)}{k},$$

where $\boldsymbol{P}^*$ is the Hilbert adjoint of $\boldsymbol{P}$.

Remark 2.6. *An equivalent definition for the spectral gap is*

$$\gamma(\boldsymbol{P}) := 1 - \sup\left(\operatorname{Spec}\left(\boldsymbol{P}|_{L_0^2(\pi)}\right)\right),$$

where $\boldsymbol{P}|_{L_0^2(\pi)}$ is the restriction of $\boldsymbol{P}$ to the closed subspace $L_0^2(\pi) := \{[f]_\pi \in L^2(\pi) : \pi(f) = 0\}$. The equivalence follows from orthogonally decomposing the Hilbert space $L^2(\pi)$ into the closed subspace $L_0^2(\pi)$ and the closed subspace of constant functions. For details, see for example Section 22.2 of Douc et al. (2018). Note further that some authors prefer defining the spectral gap using the Hilbert space $L^2(\pi, \mathbb{R})$ instead of $L^2(\pi, \mathbb{C})$, but this can easily be shown to be equivalent. We prefer to use $\mathbb{C}$, because spectral analysis is easier in complex Hilbert spaces.

Note that the above definitions depend on the choice of stationary distribution π. We omit this from notation, since it will always be clear from context what stationary distribution is meant.

The spectral gap is well-defined, since $\boldsymbol{P}$ is assumed to be self-adjoint and therefore has real spectrum. Since $\boldsymbol{P}$ has operator norm 1, it follows that $\mathrm{Spec}(\boldsymbol{P}) \subset [-1,1]$ for any reversible transition kernel P. This implies that the spectral gap is always non-negative. For the pseudo spectral gap, recall that for any $k \in \mathbb{N}$ and not-necessarily reversible transition kernel P with stationary distribution π, we have that $(\boldsymbol{P}^*)^k \boldsymbol{P}^k$ corresponds to a transition kernel with stationary distribution π. Hence the pseudo spectral gap $\gamma_{ps}(\boldsymbol{P})$ is also well-defined. In fact, $(\boldsymbol{P}^*)^k \boldsymbol{P}^k$ is positive semi-definite and hence $\mathrm{Spec}\big((\boldsymbol{P}^*)^k \boldsymbol{P}^k\big) \subset [0,1]$. It immediately follows that the pseudo spectral gap is always in $[0,1]$.

For a reversible Markov chain $(\Phi_i)_i$, Paulin (2015) uses the spectral gap to study sums of the form $\sum_{i=1}^n f(\Phi_i)$ for some function f. In particular, a larger spectral gap implies that such sums behave more like iid sums, see for example his Theorem 3.1. As an analogue to the spectral gap for non-reversible Markov chains, Paulin (2015) introduced the pseudo spectral gap. In particular, his Theorem 3.4 assumes a positive pseudo spectral gap.

To better understand the pseudo spectral gap, the following proposition offers ways to show that a Markov chain has positive pseudo spectral gap. The proof can be found in Appendix B.

Proposition 3. *Let P and Q be transition kernels on Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then both of the following hold:*

(i) If P is uniformly ergodic with stationary distribution π, then P is N(π) and has positive pseudo spectral gap $\gamma_{ps}(\boldsymbol{P}) > 0$. If P and Q are both uniformly ergodic, then $P \times Q$ is also uniformly ergodic.

(ii) If P has a stationary distribution π and $\gamma_{ps}(\boldsymbol{P}) > 0$, then $P \times P$ also has positive pseudo spectral gap and $\gamma_{ps}(\boldsymbol{P} \times \boldsymbol{P}) \geq \gamma_{ps}(\boldsymbol{P})$.

The first statement of (i) states that uniform ergodicity is sufficient for positivity of the pseudo spectral gap. To interpret the rest of Proposition 3, it is useful to note that $P \times Q$ is the transition kernel of the Markov chain $(\Phi_i, \Psi_i)_i$, if $(\Phi_i)_i$ and $(\Psi_i)_i$ are two independent Markov chains with transition kernels P and Q respectively. Then (i) states that independent uniformly ergodic Markov chains have a positive joint pseudo spectral gap. (ii) states that independent Markov chains with the same transition kernel with positive pseudo spectral gap also have a positive joint pseudo spectral gap.

2.4.2 Subgaussianity

As was the case for the iid assumption on $(\boldsymbol{X}_i)_{i=1}^n$, the proofs of Theorems $1'$ and $2'$ rely on the iid standard normality of $(\epsilon_i)_{i=1}^n$ and their independence from $(\boldsymbol{X}_i)_{i=1}^n$ only through Lemma $4'$. Furthermore, the proofs do not depend on the particular constants appearing in Lemma $4'$. Hence, the standard normality assumption on the errors can be relaxed to assuming that they follow some other distribution so that Lemma $4'$ holds, possibly with different constants. The proof of Lemma $4'$ relies on the iid standard normality assumption through the following two properties:

- the second moments $\mathbb{E}[\epsilon_i^2]$ are bounded uniformly in i,
- there exists a constant $C > 0$ such that for all n, large enough $M \in \mathbb{N}$ and functions $w_1, ..., w_M : \left([0,1]^d\right)^n \to \{\boldsymbol{w} \in \mathbb{R}^n : |\boldsymbol{w}|_2 = 1\}$, we have

$$\mathbb{E}\left[\max_{j\in[M]} \left((\epsilon_i)_{i=1}^n \cdot w_j\left((\boldsymbol{X}_i)_{i=1}^n\right)\right)^2\right] \leq C \log M,$$

  where $\cdot$ denotes the dot product.

Note that the second property follows not only from the iid standard normality, but also from $(\epsilon_i)_{i=1}^n$ being independent of $(\boldsymbol{X}_i)_{i=1}^n$.

Under appropriate conditions on the dependence between $(\epsilon_i)_{i=1}^n$ and $(\boldsymbol{X}_i)_{i=1}^n$, both of these properties can also be made to hold for subgaussian errors instead of standard normal ones. Hence, the proof of Lemma $4'$ can be easily modified to allow for subgaussian errors instead of standard normal errors. Furthermore, both the assumption of independence and the assumption of identical distributions can be relaxed. Subgaussianity is defined as follows:

Definition 6. *Let Z be a random variable taking values in $\mathbb{R}$, let $\boldsymbol{Z}$ be a random vector* *taking values in $\mathbb{R}^m$ for some $m \in \mathbb{N}$, and let $\nu > 0$. Z is ν-subgaussian if*

$$\mathbb{E}\left[\exp(tZ)\right] \le \exp(\nu^2 t^2/2) \quad \forall t \in \mathbb{R}.$$

In that case, ν^2 is called a variance proxy of Z. $\boldsymbol{Z}$ is ν-subgaussian if $\boldsymbol{a}^\top \boldsymbol{Z}$ is ν-subgaussian for each $\boldsymbol{a} \in \mathbb{R}^m$ with $|\boldsymbol{a}|_2 = 1$. $\boldsymbol{Z}$ is subgaussian if it is ν-subgaussian for some $\nu > 0$.

Note that the variance proxy for a subgaussian random vector or random variable is not unique and that if ν^2 is a variance proxy, then also λ^2 is a variance proxy for any $\lambda^2 > \nu^2$. Furthermore, it is well-known that subgaussian variables have mean zero and a finite variance bounded by their variance proxy. Common examples of subgaussian random variables are bounded random variables with mean zero and centered normal random variables. In particular, a centered normal random variable is subgaussian with its variance as variance proxy. One can also show that centered jointly Gaussian random vectors are subgaussian. Indeed, suppose $\boldsymbol{Z} \sim N(0, \Sigma)$ for some positive semi-definite matrix $\Sigma \in \mathbb{R}^{k \times k}$. Then for every vector $\boldsymbol{a} \in \mathbb{R}^k$ with $|\boldsymbol{a}|_2$, we have that $\boldsymbol{a}^\top \boldsymbol{Z} \sim N(0, \boldsymbol{a}^\top \Sigma \boldsymbol{a})$ and hence that $\boldsymbol{a}^\top \boldsymbol{Z}$ is $\boldsymbol{a}^\top \Sigma \boldsymbol{a}$-subgaussian. Now recall that it is well-known that the quadratic form $\boldsymbol{a} \mapsto \boldsymbol{a}^\top \Sigma \boldsymbol{a}$ on the unit sphere $\{\boldsymbol{a} \in \mathbb{R}^k : |\boldsymbol{a}_2| = 1\}$ has maximal value λ equal to the largest eigenvalue of Σ. Hence, $\boldsymbol{Z}$ is $\sqrt{\lambda}$-subgaussian.

To allow dependence between $(\epsilon_i)_i$ and $(\boldsymbol{X}_i)_i$, we introduce the concept of conditional subgaussianity. Before we give the definition of subgaussianity, recall that for random vectors $\boldsymbol{Z}, \boldsymbol{U}$ taking values in $\mathbb{R}^m, \mathbb{R}^k$ respectively, there exists a probability kernel between $\mathbb{R}^k$ and $\mathbb{R}^m$ that represents the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{U}$, see for example Theorem 5.3 of Kallenberg (1997) for a proof and more technical details. We can now define conditional subgaussianity using such a probability kernel.

Definition 7. *Let $\nu > 0$ and let $\boldsymbol{Z}, \boldsymbol{U}$ be random vectors taking values in $\mathbb{R}^m, \mathbb{R}^k$ respectively for some $m, k \in \mathbb{N}$. $\boldsymbol{Z}$ is conditionally ν-subgaussian given $\boldsymbol{U}$ if there exists a probability kernel P between $\mathbb{R}^m$ and $\mathbb{R}^k$ that represents the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{U}$ such that $P(\boldsymbol{u}, \cdot)$ is a ν-subgaussian distribution for each $\boldsymbol{u} \in \mathbb{R}^k$. $\boldsymbol{Z}$ is conditionally subgaussian given $\boldsymbol{U}$ if there exists a constant $\lambda > 0$ such that $\boldsymbol{Z}$ is conditionally λ-subgaussian given $\boldsymbol{U}$.*

Note that requiring conditional ν-subgaussianity of $(\epsilon_i)_i$ is stronger than only requiring ν-subgaussianity. It follows immediately from the law of iterated expectations that a conditional ν-subgaussian random vector is also unconditionally ν-subgaussian. Furthermore, the fact that subgaussian random vectors are centered, means that conditionally subgaussian random vectors are conditionally centered.

2.4.3 A generalization of Theorem 1′

Having reviewed the necessary definitions, we are now ready to state our generalization of Theorem 1′.

Proposition 4. *Let $d \in \mathbb{N}, n \in \mathbb{N}, \nu > 0$ and $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ for some $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ as in* (2)*. Let $C_{(ii)}, C_{(iii)}, C^-_{(iv)}, C^+_{(iv)} > 0$ be constants. Let $F \geq 1$, $L > 0, \boldsymbol{p} = (p_i)_{i=0}^{L+1} \in \mathbb{N}^{q+2}, s \in \mathbb{N}$, with $p_0 = d, p_{L+1} = 1$. Suppose that (i), (iii) and (iv) of Theorem 1 and (ii') of Theorem 1′ hold. Consider the d-variate nonparametric regression model* (1) *with regression function f_0, but instead of $(\boldsymbol{X}_i)_{i=1}^n$ being iid and $(\epsilon_i)_{i=1}^n$ being iid standard normal and independent of $(\boldsymbol{X}_i)_{i=1}^n$, suppose that both of the following hold:*

- *(a) $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ is a strictly stationary and $N(\pi)$ Markov chain with state space $[0,1]^d$, transition kernel P, unique stationary distribution π and positive pseudo spectral gap $\gamma_{ps}(\boldsymbol{P}) > 0$.*
- *(b) $(\epsilon_i)_{i=1}^n$ is a random vector taking values in $\mathbb{R}^n$ and conditionally ν-subgaussian given $(\boldsymbol{X}_i)_{i=1}^n$.*

Furthermore, let $\hat{f}_n$ be an estimator taking values in $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$. Then there exist constants $C, C' > 0$ depending only on q, $\boldsymbol{d}$, $\boldsymbol{t}$, $\boldsymbol{\beta}$, F, $C_{(ii)}$, $C_{(iii)}$, $C^-_{(iv)}$, $C^+_{(iv)}$, ν, $\gamma_{ps}(\boldsymbol{P})$ such that (5) *holds if $\Delta_n(\hat{f}_n, f_0) \leq C\phi_n L \log^2 n$, and* (6) *holds if $\Delta_n(\hat{f}_n, f_0) \geq C\phi_n L \log^2 n$.*

The proof can be found in Appendix B. We will now discuss this result and its implications.

We note first of all that Proposition 4 is a strict generalization of Theorem 1′. Suppose that $(\boldsymbol{X}_i)_i$ is iid with $\boldsymbol{X}_1 \sim \pi$ for some probability distribution π. It is then easy to see that $(\boldsymbol{X}_i)_i$ is an $\mathrm{N}(\pi)$ Markov chain with transition kernel $P(\boldsymbol{x}, \cdot) = \pi$ for all $\boldsymbol{x} \in [0,1]^d$. The operator $\boldsymbol{P}$ is then given by $\boldsymbol{P}[f]_\pi = [\boldsymbol{x} \mapsto \mathbb{E}[f(\boldsymbol{X}_1)]]_\pi$ for any $[f]_\pi \in L^2(\pi)$. Note in particular that $\boldsymbol{P}[f]_\pi$ is an equivalence class of π-a.e. constant functions. Furthermore, for any $[f]_\pi \in L_0^2(\pi)$, we have that $\boldsymbol{P}[f]_\pi$ is the equivalence class of the zero function. Now note that $\boldsymbol{P}$ is idempotent and self-adjoint. Hence, $\boldsymbol{P}^*\boldsymbol{P} = \boldsymbol{P}$. Using the equivalent alternative definition of the spectral gap from Remark 2.6, we then obtain

$$\gamma(\boldsymbol{P}^*\boldsymbol{P}) \quad = \quad \gamma(\boldsymbol{P}) \quad := \quad 1 - \sup\left(\operatorname{Spec}\left(\boldsymbol{P}|_{L_0^2(\pi)}\right)\right) \quad = \quad 1,$$

since the spectrum of the zero function is $\{0\}$. This implies that $\gamma_{ps}(\boldsymbol{P}) = 1$, since the pseudo spectral gap always lies in $[0,1]$. Hence, any iid sequence $(\boldsymbol{X}_i)_i$ satisfies (a) of Proposition 4 with $\gamma_{ps}(\boldsymbol{P}) = 1$. Next, suppose $(\epsilon_i)_i$ is iid standard normal. Note that the standard normal distribution is 1-subgaussian and a random vector $\boldsymbol{Z}$ of n iid standard normal random variables is also 1-subgaussian, since $\boldsymbol{a}^\top \boldsymbol{Z} \sim N(0,1)$ for any vector $\boldsymbol{a} \in \mathbb{R}^n$ with $|\boldsymbol{a}|_2$. Hence, $(\epsilon_i)_{i=1}^n$ is 1-subgaussian. If $(\epsilon_i)_{i=1}^n$ is independent of $(\boldsymbol{X}_i)_{i=1}^n$, then $(\epsilon_i)_{i=1}^n$ is also conditionally 1-subgaussian given $(\boldsymbol{X}_i)_{i=1}^n$. Hence, (b) of Proposition 4 is satisfied with $\nu = 1$.

In many statistical problems in practice, the iid assumptions of Theorem 1 are not realistic. In time series analysis, for example, the regressors at different timepoints are often dependent. It is common to model such dependence in time series using a Markov chain structure. In a similar vein, the errors in a time series context often exhibit serial dependence as well. Assuming joint subgaussianity is one way to allow for such dependence. Hence, Proposition 4 can be seen as a generalization of Theorem $1'$ especially focused on regression with time series. Although Markov chains are indeed specifically a concept geared towards time series, the conditional subgaussianity assumption can also apply in other contexts with dependent errors, such as the case of spatial data.

Markov chains have been studied extensively, but unfortunately, the pseudo spectral gap has not been studied nearly as extensively. Furthermore, it is often difficult to calculate the pseudo spectral gap for a given transition kernel. Hence, it could be difficult to determine whether a transition kernel satisfies (a) of Proposition 4. Fortunately, Proposition 3(i) implies that uniform ergodicity is a sufficient condition for (a) of Proposition 4. Uniform ergodicity is usually seen as a rather strong assumption for a Markov chain on a general state space, but the class of uniformly ergodic transition kernels on $[0,1]^d$ is still rich. To demonstrate this, consider (i) and (v) of Theorem 16.0.2 of Meyn and Tweedie (2012). That is, uniform ergodicity of a transition kernel P on $\mathcal{X}$ is equivalent to the existence of a constant m and non-zero measure μ on $\mathcal{X}$ such that $P^m(x, A) \geq \mu(A)$ for all $x \in \mathcal{X}$ and Borel sets $A \subset \mathcal{X}$. Hence, P is uniformly ergodic if there exist a constant $\kappa > 0$ and finite measure μ on $\mathcal{X}$ such that for all $x \in \mathcal{X}$, we have that $P(x, \cdot)$ has a μ-density bounded from below by κ. In our particular case, $\mathcal{X} = [0,1]^d$, so that we can choose Lebesgue measure as our finite measure μ. Therefore, (a) of Proposition 4 is satisfied for any transition kernel with a constant $\kappa > 0$ such that for each $\boldsymbol{x} \in [0,1]^d$, $P(\boldsymbol{x}, \cdot)$ has a Lebesgue density bounded from below by κ. This condition is usually much easier to verify for transition kernels on $[0,1]^d$ than calculating the pseudo spectral gap directly. As an explicit example of a class of transition kernels satisfying this condition, consider any positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$ and define a transition kernel on $[0,1]^d$ by

$$P_\Sigma(\boldsymbol{x}, A) := \mathbb{P}_{\boldsymbol{Z} \sim N(\boldsymbol{x}, \Sigma)} \left(\boldsymbol{Z} \in A | \boldsymbol{Z} \in [0,1]^d \right).$$

A similar construction can be used with other families of distributions for $\boldsymbol{Z}$ instead of the Gaussian distributions.

As an alternative to the pseudo spectral gap, a recent preprint on arXiv by Jiang et al. (2020) considers the right spectral gap $\gamma((\boldsymbol{P} + \boldsymbol{P}^*)/2)$. Their Theorem 2 is a generalization of Bernstein's inequality for Markov chains with positive right spectral gap and, if the theorem is true, could be used to replace the pseudo spectral gap in (a) of Proposition 4 by the right spectral gap.

Note that the conditional subgaussianity assumption not only allows dependence between the errors of different observations, but also dependence between the errors and the regressors. In fact, even dependence between the error ϵ_i and regressors $\boldsymbol{X}_j$ of different observations $i \neq j$ is possible. However, the dependence allowed by conditional subgaussianity is limited by the fact that it requires conditional mean zero $\mathbb{E}\left[(\epsilon_i)_{i=1}^n | (\boldsymbol{X}_i)_{i=1}^n\right] = 0$.

Still, an important type of dependence that is possible under the conditions of Proposition 4 is conditional heteroskedasticity. That is, $\mathbb{E}[\epsilon_i^2|(\boldsymbol{X}_j)_{j=1}^n]$ might be a non-constant function of $(\boldsymbol{X}_j)_{j=1}^n$ for each i, and $\mathbb{E}[\epsilon_i^2]$ might be different for different i. Note that conditional ν-subgaussianity does imply that $\mathbb{E}[\epsilon_i^2|(\boldsymbol{X}_j)_{j=1}^n]$ must be bounded from above by ν^2 for each i, and hence also that $\mathbb{E}[\epsilon_i^2] \leq \nu^2$ for all i. Still, (b) of Proposition 4 allows for considerable flexibility in the joint distribution of $(\boldsymbol{X}_i, \epsilon_i)_{i=1}^n$. As an example to showcase this flexibility, suppose that $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ satisfies (a) of Proposition 4 and recall that for any positive semi-definite real matrix Σ, the $N(0,\Sigma)$ distribution is subgaussian with as variance proxy the largest eigenvalue $\lambda_1(\Sigma)$. Now suppose that for any n, we have a measurable function

$$\Sigma_n : ([0,1]^d)^n \to \mathcal{M}_n(\nu) := \{\Sigma \in \mathbb{R}^{n\times n} : \Sigma \text{ is positive semi-definite and } \lambda_1(\Sigma) \leq \nu^2\}.$$

Then suppose $(\epsilon_i)_{i=1}^n|(\boldsymbol{X}_i)_{i=1}^n \sim N\left(0, \Sigma_n\left((\boldsymbol{X}_i)_{i=1}^n\right)\right)$. This implies that $(\epsilon_i)_{i=1}^n$ is conditionally ν-subgaussian given $(\boldsymbol{X}_i)_{i=1}^n$, and hence $(\boldsymbol{X}_i, \epsilon_i)_{i=1}^n$ satisfies the requirements of Proposition 4. Note that $\mathcal{M}_n(\nu)$ contains a covariance matrix for any possible correlation structure, since the zero matrix is in $\mathcal{M}_n(\nu)$ and any non-zero positive semi-definite matrix $\Sigma \in \mathbb{R}^{n\times n}$ can be multiplied by $\nu^2/\lambda_1(\Sigma)$ to obtain a matrix in $\mathcal{M}_n(\nu)$. We emphasise that we require no further conditions on the function Σ_n other than measurability. Furthermore, if $(\boldsymbol{X}_i)_i$ is not almost surely constant, we can very easily choose a function Σ_n such that $(\epsilon_i)_{i=1}^n$ is not covariance stationary nor unconditionally jointly Gaussian, and such that $(\epsilon_i)_{i=1}^n$ and $(\boldsymbol{X}_i)_{i=1}^n$ are dependent with both conditional and unconditional heteroskedasticity.

As a more explicit example of an interesting error sequence, suppose $a > 0$ and $\rho \in [0,1)$ are constants and suppose $(u_i)_{i=1}^\infty$ is a sequence of iid $N(0,a^2)$ random variables independent of $(\boldsymbol{X}_i)_{i=1}^\infty$. Define $(\epsilon_i)_{i=1}^\infty$ by

$$\epsilon_{i+1} = \rho\epsilon_i + u_i \text{ for all } i \in \mathbb{N}, \quad \epsilon_1 \sim N\left(0, a^2/(1-\rho^2)\right).$$

That is, $(\epsilon_i)_{i=1}^\infty$ is a centered stationary Gaussian AR(1) process independent of $(\boldsymbol{X}_i)_{i=1}^\infty$.

Then it is well-known that $(\epsilon_i)_{i=1}^n \sim N(0, a^2 K_n(\rho))$, where $K_n(\rho)$ is the $(n \times n)$ symmetric Toeplitz matrix with (i,j)-th entry $\rho^{|i-j|}$, under the convention that $0^0 = 1$. Such matrices are called Kac-Murdock-Szegö matrices and it is well-known that the largest eigenvalue of $K_n(\rho)$ is bounded by $(1-\rho^2)/(1-2\rho+\rho^2)$, see for example Example (c) in Section 5.3 of Grenander and Szegö (1958). In particular, if we choose $\nu^2 = a^2(1-\rho^2)/(1-2\rho+\rho^2)$, then $(\epsilon_i)_{i=1}^n$ satisfies (b) of Proposition 4 for each n.

3 Examples and comparisons

In this section, we discuss some specific examples of regression functions and compare the rate guaranteed by Theorem $1'$ with rates obtained in related literature by other estimators or in other settings. In each of these example we will consider the d-variate nonparametric regression model (1). For each n, we will further consider an NN-valued estimator $\hat{f}_n$ taking values in $\mathcal{F}_0(L_n, \boldsymbol{p}^{(n)}, s_n, F)$ for some appropriate hyperparameters $L_n, \boldsymbol{p}^{(n)}, s_n, F$ that are to be specified for each example specifically.

3.1 Non-composite functions

Suppose $\beta > 0$ and that f_0 is β-Hölder. Then for K large enough,

$$f_0 \in \mathcal{C}_d^\beta([0,1]^d, K) = \mathcal{G}(q, d, t, \beta, K),$$

where $q = 0$ and $t = d$. We then have that $\phi_n = n^{-\beta/(2\beta+d)}$ in this context. Now suppose that F is sufficiently large and that the network hyperparameters $L_n, \boldsymbol{p}^{(n)}, s_n$ grow at an appropriate rate satisfying the requirements of Theorem $1'$ with in particular $L_n \asymp \log n$. If $\hat{f}_n$ is an empirical risk minimizer for the class $\mathcal{F}_0(L_n, \boldsymbol{p}^{(n)}, s_n, F)$, then Corollary 1 guarantees that

$$R(\hat{f}_n, f_0) \lesssim n^{-\frac{2\beta}{2\beta+d}} \log^3 n.$$

In fact, Theorem $1'$ implies that this is also true if $(\hat{f}_n)_n$ are only approximate empirical risk minimizers, in the sense that $\Delta_n(\hat{f}_n, f_0)/(\phi_n \log^3 n) \overset{n\to\infty}{\longrightarrow} 0$. If we further assume that K is large enough and that the distribution of $\boldsymbol{X}_1$ has Lebesgue density bounded from above and from below by positive constants, then Theorem 3 implies that the minimax rate for the expected squared prediction error for estimation of regression functions in $\mathcal{C}_d^\beta([0,1]^d, K)$ is at least $n^{-2\beta/(2\beta+d)}$. In particular, $(\hat{f}_n)_n$ then achieves this rate up to $\log n$ factors. Note that the rate $n^{-2\beta/(2\beta+d)}$ corresponds to the minimax L^2-rate found by Stone (1982) under similar but slightly different assumptions.

3.2 Varying coefficient models

Suppose that $d = k + m$ for some $k, m \in \mathbb{N}$ and partition any vector $\boldsymbol{x} \in [0,1]^d$ into $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{w})$ with $\boldsymbol{u} \in [0,1]^k, \boldsymbol{w} \in [0,1]^m$. For each i, partition $\boldsymbol{X}_i = (\boldsymbol{U}_i, \boldsymbol{W}_i)$ in the same way. Suppose further that $\rho, C > 0$ are constants and that f_0 is of the form $f_0(\boldsymbol{x}) = \boldsymbol{w}^\top \theta(\boldsymbol{u})$, for some bounded ρ-Hölder function $\theta = (\theta_j)_{j=1}^k : [0,1]^k \to [-C, C]^m$. This is a generalization of the varying coefficient model introduced by Hastie and Tibshirani (1993). We can then write $f_0 = g_2 \circ g_1 \circ g_0$, where

$$\begin{aligned} g_0 : [0,1]^d \to [-C,C]^m \times [0,1]^m, &\quad g_0(\boldsymbol{x}) := (\theta(\boldsymbol{u}), \boldsymbol{w}) \\ g_1 : [-C,C]^m \times [0,1]^m \to [-C,C]^m, &\quad g_1(\boldsymbol{b}, \boldsymbol{w}) := (b_1 w_1, ..., b_m w_m) \\ g_2 : [-C,C]^m \to [-mC, mC]^m, &\quad g_2(\boldsymbol{z}) = \sum_{j=1}^m z_j. \end{aligned}$$

Note that the Hölder smoothness of g_0 is equal to that of θ, and that g_1 and g_2 are both infinitely differentiable and hence have arbitrarily high Hölder smoothness. Furthermore, if $g_0 = (g_{0j})_{j=1}^{2m}$, then g_{0j} depends on only k of its input variables if $j \leq m$ and on only 1 if $j > m$. If $g_1 = (g_{1j})_{j=1}^m$, then each g_{1j} depends on only two of its $2m$ input variables. Finally, g_2 depends on all m of its input variables. Hence, $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$, where $q = 2$, $\boldsymbol{d} = (d, 2m, m, 1)$, $t = (k, 2, m)$, $\beta_0 = \rho$, $\beta_1 > 1$, $\beta_2 > 1$, and K is sufficiently large. Then $\beta_i^* = \beta_i$ for all i. Since β_1, β_2 can be chosen arbitrarily large, they can in particular be chosen so that

$$\phi_n := \max\left\{ n^{\frac{-2\rho}{2\rho+k}},\ n^{\frac{-2\beta_1}{2\beta_1+2}},\ n^{\frac{-2\beta_2}{2\beta_2+m}} \right\} = n^{\frac{-2\rho}{2\rho+k}}.$$

Now suppose that $F, L_n, \boldsymbol{p}^{(n)}, s_n$ are chosen appropriately for Theorem $1'$ with $L_n \asymp \log n$, as we did in the previous example. Then we obtain

$$R(\hat{f}_n, f_0) \lesssim n^{\frac{-2\rho}{2\rho+k}} \log^3 n.$$

In particular, this rate depends only on the Hölder smoothness ρ and input dimension k of θ, and it does not depend on d except through k. Hence, the rate $n^{\frac{-2\rho}{2\rho+k}} \log^3 n$ can be achieved for any m. This is an example of how the curse of dimensionality can be circumvented in nonparametric regression. In particular, the estimators $(\hat{f}_n)_n$ seem to be able to take advantage of the structure in the regression function. We emphasise, however, that we did use knowledge of this structure in choosing the network hyperparameters that lead to this estimation rate. Intuitively, one might view this structural assumption on the regression function as isolating the "nonparametric dependence" of Y_i on $\boldsymbol{X}_i$ to the k variables in $\boldsymbol{U}_i$. Indeed, the rate $\phi_n \log^3 n$ in this case agrees, up to $\log n$ factors, with the minimax L^2-rate given by Stone (1982) for estimation of ρ-Hölder regression functions of k variables.

In practice, θ in the varying coefficient model is often estimated using local linear estimation, see for example the description by Park et al. (2015) for $k = 1$. A local linear estimator $\tilde{\theta}^{(n)}$ of θ can be used to obtain an estimator $\tilde{f}_n$ of f_0 by defining $\tilde{f}_n(\boldsymbol{x}) = \boldsymbol{w}^\top \tilde{\theta}^{(n)}(\boldsymbol{u})$. Note that the local linear estimator very explicitly relies on the structure of the regression function. The pointwise asymptotic behaviour for a local linear estimator $\tilde{\theta}^{(n)}(\boldsymbol{u}_0)$ of $\theta(\boldsymbol{u}_0)$ for some fixed interior point $\boldsymbol{u}_0 \in (0,1)^k$ is well-known. In particular, if $k = 1$ and θ is twice continuously differentiable, then an appropriate choice of bandwidth can lead to an $n^{-4/5}$-rate for the mean squared error, see for example Theorem 1 of Wong et al. (2008), where the varying coefficient model is a special case of their more general model. This corresponds to the rate ϕ_n with $k = 1$ and $\rho = 2$.

A similar approach can be used to find fast estimation rates by NN-valued estimators for other popular nonparametric regression models that make structural assumptions on the regression function. Schmidt-Hieber (2020), for example, discusses the generalized additive model and regression functions with a product structure.

3.3 Neural networks as regression functions

Schmidt-Hieber (2020) considered the class $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ because NNs can also be seen as compositions of functions and might therefore be well-suited for estimation of other composite functions. Continuing this philosophy, one might wonder how well NN-valued estimators can estimate a regression function that is also an NN. Suppose that $f_0 \in \mathcal{F}_0(L_0, \boldsymbol{p}^{(0)})$ for some fixed network architecture $(L_0, \boldsymbol{p}^{(0)})$ and consider a representation as in (3) for f_0. This gives a natural way to view f_0 as the composition of the $2L+1$ functions $\sigma_{\boldsymbol{v}_i}$ and W_i. Note that W_i is linear and therefore has an arbitrarily large Hölder smoothness $\rho_i > 1$. On the other hand, $\sigma_{\boldsymbol{v}_i} = (\sigma_{v_{ij}})_{j=1}^{p_i^{(0)}}$, and each $\sigma_{v_{ij}}$ is 1-Hölder and depends only on one of the $p_i^{(0)}$ input variables. Furthermore, note that the domain of f_0 is bounded and all network parameters are bounded in absolute value by 1. Since the network architecture is fixed, this means that we can consider all functions $\sigma_{\boldsymbol{v}_i}$ and W_i as bounded functions on bounded domains. Hence, $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$, where we set $q = 2L+1$, $\boldsymbol{d} = (d, p_1^{(0)}, p_1^{(0)}, p_2^{(0)}, p_2^{(0)}, ..., p_L^{(0)}, p_L^{(0)}, 1)$, $\boldsymbol{t} = (d, 1, p_1^{(0)}, 1, p^{(0)})_2, ..., 1, p_L^{(0)})$, $\boldsymbol{\beta} = (\rho_0, 1, \rho_1, 1, ..., 1, \rho_L)$ and some sufficiently large K. Note that since $\beta_i \geq 1$ for all i, we again have $\beta_i^* = \beta_i$ for all i. Now let $\rho > 1$ and set $\rho_i = \rho$ for all i. We then have

$$\frac{2\beta_i^*}{2\beta_i^* + t_i} = \begin{cases} \frac{2\rho}{2\rho + p_k^{(0)}} & \text{if } i \text{ is even with } i = 2k \\ \frac{2}{3} & \text{otherwise.} \end{cases}$$

In particular, ρ can be chosen large enough so that $2\rho/(2\rho + p_k^{(0)}) > 2/3$ for all k. In that case, we have $\phi_n = n^{-2/3}$. Choosing an appropriate sequence of network hyperparameters $L_n, \boldsymbol{p}^{(n)}, s_n$ and F large enough, then ensures by Theorem 1′ that the estimators $(\hat{f}_n)_n$ can estimate f_0 with $R(\hat{f}_n, f_0) \lesssim n^{-2/3} \log^3 n$. Note that this is again independent of d.

Since we assumed a fixed network architecture, this is a parametric regression model and the rate being independent of d is not surprising. In fact, from that perspective, the rate $n^{-2/3} \log^3 n$ is slow compared to the n^{-1} rate for the mean squared prediction error that is usually achieved for parametric models. However, the estimators $(\hat{f}_n)_n$ are more flexible than parametric estimators, since the estimators $(\hat{f}_n)_n$ are able to estimate any function in the infinite-dimensional space $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ with the rate $n^{-2/3} \log^3 n$. On the other hand, parametric estimators with a fixed number of parameters can only achieve the rate n^{-1} on a strict subset of $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$. Furthermore, we emphasise that this example does not prove that NN-valued estimators cannot achieve the rate n^{-1} when estimating NNs, since we only considered NN-valued estimators that satisfy the requirements of Theorem $1'$. Theorem $1'$ can never guarantee a rate of n^{-1}, since ϕ_n is of the form $n^{-2b/(2b+t)}$ for some $b > 0$ and $t \in \mathbb{N}$, and hence $\phi_n = n^{-a}$ for some $a \in (0, 1)$.

4 Practical issues

In this section, we discuss some points related to translating the results of Schmidt-Hieber (2020) to statistical analysis in practice. We will discuss assumptions on the data, the choice of hyperparameters $(L, \boldsymbol{p}, s, F)$, and the training of sparse NNs.

4.1 Assumptions on the data

The nonparametric regression model (1) should be seen as an assumption on the joint distribution of the sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$. In practice, one can often justify the boundedness and iid assumptions on $(\boldsymbol{X}_i)_i$ based on how the data were obtained. Data from experiments performed in controlled settings can often be assumed to be iid, for example. The iid standard normality assumption on the errors is not as realistic, but can be relaxed considerably, as we showed in Proposition 4. However, we suspect that it will usually be hard to justify the assumption of Theorem 1 that $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ for some particular values of $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$. Indeed, Schmidt-Hieber (2020) did not choose this class because it appeared naturally in applications, but because NNs are also compositions of functions. Still, in some applications, one might argue as we did for the varying coefficient model in Section 3.2 to conclude that $f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ for some values of $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ that lead to Theorem $1'$ guaranteeing good performance.

4.2 Hyperparameter tuning

Theorem $1'$ assumes non-random network hyperparameters $L, \boldsymbol{p}, s$, which may vary with n. Although (ii'), (iii) and (iv) of Theorems $1'$ and 1 provide bounds on the asymptotic orders of these network hyperparameters, such asymptotic bounds offer little to no guidance for choosing a network architecture for a particular sample size n. Furthermore, these asymptotic bounds depend on $q, \boldsymbol{t}$ and $\boldsymbol{\beta}$, which will often be unknown. Similarly, Theorem $1'$ suggests that the hyperparameter F should be chosen based on K, but K will usually be unknown as well. Hence, in practice, network hyperparameters are often chosen based on data-based methods, such as cross-validation, bootstrap-based procedures, or even "manual" tuning. See for example the overview by Yang and Shami (2020) on hyperparameter optimization. Using data-based methods makes the hyperparameters random, which means that Theorem $1'$ can no longer be applied. However, data-based hyperparameters could potentially improve the estimation performance compared to fixed non-random hyperparameters, since data-based methods might be able to adapt to the particular realization of the sample. This could be an interesting topic for future research.

4.3 Training sparse neural networks

Theorem $1'$ and Corollary 1 suggest that it is desirable for an $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$-valued estimator to be close to an empirical risk minimizer, if the hyperparameters $L, \boldsymbol{p}, s, F$ are chosen appropriately. In practice, both sparse and non-sparse NNs are usually fit by trying to minimize empirical risk using methods based on stochastic gradient descent. However, guaranteeing true empirical risk minimization for large samples and large NN architectures is not always feasible. The main challenges are that the empirical risk is a non-convex function of a large number of parameters, and that there are usually many suboptimal local minima to get stuck in. This so-called loss landscape is a very active research topic. See for example the recent work by He et al. (2019), or the work by Liu et al. (2021) based on it. These papers show the existence of suboptimal local minima and saddle points in the loss landscape of non-sparse ReLU NNs. The extension of these results to the class $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$ might be an interesting topic for future research.

The fitting of sparse NNs specifically has also received much recent attention in the literature, mainly motivated by computational reasons. A popular general approach to training sparse networks is to first train a non-sparse network and then "pruning", that is, setting certain parameters to 0. This idea is not new and was, for example, already the approach taken by Le Cun et al. (1989) in their Optimal Brain Damage method. However, the downside to such methods is that one must first train a non-sparse network, which is computationally expensive. An alternative approach is to incorporate sparsity in the training procedure from the start, instead of first training a non-sparse network. This generally has the advantage of requiring less training time. See for example the RigL method introduced by Evci et al. (2020).

However, most of the methods in the literature seem based on heuristics or experimental evidence. Furthermore, their study of sparsity is usually motivated by computational reasons and the goal is to obtain an NN that can be run faster with the same accuracy as a non-sparse network. Schmidt-Hieber (2020), on the other hand, uses sparsity as an explicit form of regularization to prevent overfitting and improve the accuracy of the resulting NN compared to a non-sparse one. It might be an interesting topic for future research to analyze the behaviour of $\Delta_n(\hat{f}_n, f_0)$ for methods such as Optimal Brain Damage (Le Cun et al., 1989), or RigL (Evci et al., 2020).

5 Conclusion

In this thesis, we have studied the work of Schmidt-Hieber (2020) on estimation of a nonparametric regression model using ReLU NN-valued estimators. In particular, Schmidt-Hieber (2020) proposes explicit regularization of the estimators by considering sparse NNs with parameters bounded in absolute value by 1. We have seen that regression functions with certain composition structures can be estimated with fast rates by certain ReLU NN-valued estimators, if the network architectures and sparsity levels are chosen appropriately. In fact, we have seen that such estimators can achieve the minimax rate up to $\log n$ factors for some of these classes of regression functions. Furthermore, depending on the particular composition structure, this rate need not suffer from the curse of dimensionality. Including irrelevant regressors, for example, was shown to affect the guaranteed rate of estimation only by a constant factor. Therefore, the results of Schmidt-Hieber (2020) might offer us a way to better understand why NNs seem to perform well in practice for regression problems with many regressors. Furthermore, these results show that regularization by sparsity and bounding of parameters can lead to estimators with desirable theoretical properties. It could be interesting to study how well this approach performs in practice.

We have pointed out some points in the proofs by Schmidt-Hieber (2020) that were not completely clear to us, and we have proposed alternative approaches to solve these problems with little to no modification required to the results of Schmidt-Hieber (2020). Furthermore, a close inspection of the proofs by Schmidt-Hieber (2020) revealed that his Theorem 1 on estimation by sparse NN-valued estimators can be generalized to allow for certain Markov chains as regressors and for the errors to be conditionally subgaussian given the regressors. In particular, our generalization allows for both conditional and unconditional heteroskedasticity of the errors, and for dependence between different observations. Due to the Markov chain assumption on the regressors, our result seems especially relevant for time series analysis, but the dependence among errors and conditional heteroskedasticity could also be relevant to other fields with dependent data, such as spatial statistics. It could be an interesting topic for future research to investigate generalizations to other forms of dependence than Markov chains and conditional subgaussianity. One might, for example, consider mixing conditions instead of the Markov chain condition, and other conditional moment conditions instead of conditional subgaussianity. It might also be interesting to study whether the results can be generalized to unbounded regressors, instead of regressors in $[0,1]^d$.

We have also highlighted some practical issues related to translating the discussed theoretical results to real-life applications. We have discussed that it might be hard to justify the assumption of a certain composition structure for the regression function, but that other structural assumptions can sometimes imply a certain composition structure. We have also emphasised that the work of Schmidt-Hieber (2020) assumes non-random network hyperparameters, while these should likely be chosen in a data-based way in practice. A generalization of the results of Schmidt-Hieber (2020) to allow for data-based hyperparameters could be an interesting topic for future study. We have also discussed how one might train a sparse NN, but we observe that there is unfortunately a lack of good theoretical guarantees for the performance of currently available training algorithms. This could be an interesting subject for future research as well.

We now discuss some more possible directions for future research, apart from the ones we already mentioned before. One interesting question is whether the $\log n$ factors in the rate from Theorem 1 can be removed. Schmidt-Hieber (2020) calls these factors "likely an artifact of the proof", but gives no further argument for how to remove them. Another potentially interesting research direction is the translation of the results of Schmidt-Hieber (2020) to a classification context. Since most NNs in practice are applied to classification problems rather than regression, this seems especially relevant to us. In a similar vein, it might be interesting to study how the results of Schmidt-Hieber (2020) could be translated to NNs with activation functions other than the ReLU function. The work of Bauer and Kohler (2019) gives a partial answer to this already, but the classes of regression functions they consider are in some ways not as flexible as those of Schmidt-Hieber (2020). Lastly, we note that the results of Schmidt-Hieber (2020) rely crucially on the explicit regularization through sparsity and bounded network parameters. However, there is an active discussion in the literature about whether such explicit regularization is truly necessary for effective deep learning. Instead, some argue that the optimization method used to train the NN can already implicitly regularize and that this is enough to prevent overfitting. In particular, the idea is that the optimization method is able to distinguish the particular optima that lead to less or no overfitting. Gunasekar et al. (2018), for example, aim to make this idea exact. This leads to the question whether the explicit regularization is really necessary for the results of Schmidt-Hieber (2020), or that similar results also hold if the properties of the optimization method are taken into account.

Acknowledgements

I would like to thank Eduard Belitser for being my supervisor. I would also like to thank Harry van Zanten for initially being my supervisor, and for helping me choose a topic. Last, but certainly not least, I would like to thank Paulo Serra for being my second reader.

Appendices

A Proofs of the main results of Schmidt-Hieber (2020)

In this section, we will give detailed versions of the proofs given by Schmidt-Hieber (2020) for his main results: Theorems 1-3 and Lemma 1. At points where we find the proof given by Schmidt-Hieber (2020) unclear, we provide small modifications to the proof strategy, which will sometimes lead to slightly different results, as indicated in Section 2. In particular, we obtain slightly modified results for Theorems 1 and 2, and for Lemmas 3 and 4. We call these slightly modified results Theorems $1'$ and $2'$, and Lemmas $3'$ and $4'$ respectively. Again, all theorems and lemmas we present have been taken from Schmidt-Hieber (2020). We present the proofs in a different order than Schmidt-Hieber (2020), but we do follow his numbering of his theorems and lemmas. This means that we will present some results in an order different from their numbering. We will begin by proving Lemmas C.1, $4'$ and 5, after which we apply these to prove Theorem $2'$. Next, we prove Lemma $3'$ and apply this together with Theorem $2'$ to prove Theorem $1'$. Then we prove Theorem 3 and apply this theorem together with Theorem $2'$ to prove Lemma 1.

The only result from Schmidt-Hieber (2020) that we will use without proof is the following theorem on approximation of β-Hölder functions by sparse NNs.

Theorem 5. *For any $\beta, K > 0$, any $r, m \in \mathbb{N}$, any function $f \in \mathcal{C}_r^\beta([0,1]^r, K)$, and any integer $N \geq (\beta+1)^r \vee (K+1)e^r$, there exists a network*

$$\tilde{f} \in \mathcal{F}_0\left(L, (r, 6(r+\lceil\beta\rceil)N, ..., 6(r+\lceil\beta\rceil)N, 1\right), s, \infty\right),$$

with depth

$$L = 8 + (m+5)(1 + \lceil\log_2(r \vee \beta)\rceil),$$

and number of parameters

$$s \leq 141(r+\beta+1)^{3+r}N(m+6),$$

such that

$$||\tilde{f} - f||_{L^\infty([0,1]^r)} \leq (2K+1)(1+r^2+\beta^2)6^r N 2^{-m} + K3^\beta N^{-\beta/r}.$$

The proof of this theorem is long and tedious and can be found in Appendices A and B of the supplement to Schmidt-Hieber (2020). These appendices explicitly describe how to construct a network $\tilde{f}$ to approximate the function f. The idea of the construction is relatively straightforward. Schmidt-Hieber (2020) first approximates $x \mapsto x(1-x)$ on $[0,1]$ and uses this to approximate the multiplication $(x,y) \mapsto xy$ on $[0,1]^2$. He then uses this to approximate the multiplication $(x_i)_{i=1}^r \mapsto \prod_{i=1}^r x_i$ on $[0,1]^r$. This can then be used to approximate multivariate monomials on $[0,1]^r$, which can be used to approximate f by a local Taylor approximation.

We will now present and prove Lemma C.1, which Schmidt-Hieber (2020) uses to prove Lemma 4.

Lemma C.1. *Let* $(\eta_j)_{j=1}^M$ *be standard normal random variables. Then* $\mathbb{E}[\max_{j\in[M]} \eta_j^2] \leq 3\log M + 1$.

Proof. Let $B := \max_{j\in[M]} \eta_j^2$ and note that $B \leq \sum_{j=1}^M \eta_j^2$. Hence

$$\mathbb{E}[B] \leq M\mathbb{E}[\eta_1^2] = M.$$

Now note that for $M \in [3]$, it follows from direct calculation that $M \leq 3\log M + 1$. Hence, the claim holds for $M \in [3]$.

Now suppose $M \geq 4$. Note that for a standard normal random variable Z. we have the following bound for all $t > 0$:

$$\begin{aligned}\mathbb{P}(Z \geq t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)du \\ &\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{u}{t} \exp(-u^2/2)du \\ &= \frac{1}{\sqrt{2\pi}t} \exp(-t^2/2).\end{aligned}$$

Then note that

$$\mathbb{P}(Z^2 \geq t) \;=\; \mathbb{P}(Z \geq \sqrt{t}) + \mathbb{P}(Z \leq -\sqrt{t}) \;=\; 2\mathbb{P}(Z \geq \sqrt{t}) \;\leq\; \frac{2}{\sqrt{2\pi t}} \exp(-t/2),$$

and by the union bound then

$$\mathbb{P}(B \geq t) \;=\; \mathbb{P}\left(\bigcup_{j=1}^{M} \{\eta_j^2 \geq t\}\right) \;\leq\; \sum_{j=1}^{M} \mathbb{P}(\eta_j^2 \geq t) \;\leq\; \frac{2M}{\sqrt{2\pi t}} \exp(-t/2)$$

Hence, for any $T > 0$, the above implies that

$$\begin{aligned}
\mathbb{E}[B] &= \int_0^\infty \mathbb{P}(B \geq t)dt \\
&= \int_0^T \mathbb{P}(B \geq t)dt + \int_T^\infty \mathbb{P}(B \geq t)dt \\
&\leq T + \int_T^\infty \frac{2M}{\sqrt{2\pi t}} \exp(-t/2)dt \\
&\leq T + \frac{2M}{\sqrt{2\pi T}} \int_T^\infty \exp(-t/2)dt \\
&= T + \frac{4M}{\sqrt{2\pi T}} \exp(-T/2).
\end{aligned}$$

Choosing $T = 2\log M$ gives

$$\mathbb{E}[B] \leq 2\log M + \frac{4M}{\sqrt{4\pi \log M}} \exp(-\log M) = 2\log M + \frac{4}{\sqrt{4\pi \log M}} \leq 2\log M + 1,$$

since $M \geq 4$ implies $\sqrt{4\pi \log M} \geq 4$. This completes the proof. □

For a set of functions $\mathcal{F} \subset \{f : [0,1]^d \to \mathbb{R}\}$ and constant $\delta > 0$, denote by $\mathcal{N}(\delta, \mathcal{F}, ||\cdot||_\infty)$ the δ-covering number of $\mathcal{F}$. That is, the minimal number of closed $||\cdot||_\infty$-balls with radius δ that covers $\mathcal{F}$. Note that we do not require the centers of these balls to be in $\mathcal{F}$.

We now present Lemma 4 of Schmidt-Hieber (2020), which is basically a more general version of Theorem 2 that considers an arbitrary, sufficiently rich and uniformly bounded function class $\mathcal{F}$ rather than $\mathcal{F}(L, \boldsymbol{p}, s, F)$ specifically.

Lemma 4. *Consider the d-variate nonparametric regression model* (1) *with unknown regression function f_0. Let $\mathcal{F}$ be a set of functions and assume $\{f_0\} \cup \mathcal{F} \subset \{f : [0,1]^d \to$*

$[-F,F]\}$ *for some* $F \geq 1$. *Let* $\hat{f}$ *be an estimator taking values in* $\mathcal{F}$. *Define*

$$\Delta_n := \Delta_n(\hat{f}, f_0, \mathcal{F}) := \mathbb{E}_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(\boldsymbol{X}_i))^2 - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2\right].$$

Then for any $\epsilon, \delta \in (0,1]$ *with* $\mathcal{N}_n := \mathcal{N}(\delta, \mathcal{F}, ||\cdot||_\infty) \geq 3$, *it holds that*

$$\begin{aligned}(1-\epsilon)^2\Delta_n - F^2\frac{18\log\mathcal{N}_n + 76}{n\epsilon} - 38\delta F \\ \leq R(\hat{f}, f_0) \leq (1+\epsilon)^2\left[\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + F^2\frac{18\log\mathcal{N}_n + 72}{n\epsilon} + 32\delta F + \Delta_n\right].\end{aligned} \tag{8}$$

Although Schmidt-Hieber (2020) omits this, some measurability assumptions are necessary for f_0 and $\mathcal{F}$. In particular, the statement requires f_0 and all $f \in \mathcal{F}$ to be measurable, so that $R(\hat{f}, f_0)$ and $\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2]$ are defined. Furthermore, we were unable to follow the reasoning of Schmidt-Hieber (2020) in one step of his proof of Lemma 4. To solve this issue, we take a slightly different approach, leading to different bounds. Combining these two points, we obtain the following result:

Lemma 4′. *Lemma 4 holds if* f_0 *and all* $f \in \mathcal{F}$ *are measurable and if* (8) *is replaced by*

$$\begin{aligned}(1-\epsilon)^2\Delta_n - F^4\frac{18\log\mathcal{N}_n + 44}{n\epsilon} - 37\delta F^2 \\ \leq R(\hat{f}, f_0) \leq (1+\epsilon)^2\left(\inf_{f\in\mathcal{F}}\mathbb{E}[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + F^4\frac{18\log\mathcal{N}_n + 40}{n\epsilon} + 31\delta F^2\right).\end{aligned} \tag{9}$$

Proof. Denote $\mathbb{E} := \mathbb{E}_{f_0}$. For any function $g : [0,1]^d \to \mathbb{R}$, define $||g||_n^2 := \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i)^2$. For any estimator $\tilde{f}$ taking values in $\{\big(f : [0,1]^d \to \mathbb{R}\big) : f \text{ measurable }\}$, define $\hat{R}_n(\tilde{f}, f_0) := \mathbb{E}[||\tilde{f} - f_0||_n^2]$.

Suppose first that $\log\mathcal{N}_n \geq n$. Since $||\hat{f}||_\infty, ||f_0||_\infty \leq F$, we have that

$$\begin{aligned}R(\hat{f}, f_0) &= \mathbb{E}[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] \\ &= \mathbb{E}[\hat{f}(\boldsymbol{X})^2 + f_0(\boldsymbol{X})^2 - 2\hat{f}(\boldsymbol{X})f_0(\boldsymbol{X})] \\ &\leq 4F^2.\end{aligned}$$

In particular,

$$R(\hat{f}, f_0) \quad \leq \quad 4F^2 \quad \leq \quad (1+\epsilon)^2 F^2 \frac{18n}{n\epsilon} \quad \leq \quad (1+\epsilon)^2 F^2 \frac{18 \log \mathcal{N}_n}{n\epsilon},$$

where the final inequality follows from $\log \mathcal{N}_n \geq n$. Note that $F^4 \geq F^2$. Since all other terms in the upper bounds of Lemmas 4 and 4′ are non-negative, we have now shown that these upper bounds hold. Now let $\kappa > 0$ and suppose $\tilde{f}$ is an estimator taking values in $\mathcal{F}$ with $\Delta_n(\tilde{f}, f_0, \mathcal{F}) < \kappa$. Then

$$\begin{aligned}
\hat{R}_n(\hat{f}, f_0) - \hat{R}_n(\tilde{f}, f_0) =& \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n (\hat{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))^2 - \sum_{i=1}^n (\tilde{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))^2\right] \\
=& \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n (\hat{f}(\boldsymbol{X}_i) - (f_0(\boldsymbol{X}_i) + \epsilon_i))^2 - \sum_{i=1}^n (\tilde{f}(\boldsymbol{X}_i) - (f_0(\boldsymbol{X}_i) + \epsilon_i))^2\right] \\
&+ \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n 2\epsilon_i(\hat{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i)) - \epsilon_i^2 - 2\epsilon_i(\tilde{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i)) + \epsilon_i^2\right] \\
=& \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n (\hat{f}(\boldsymbol{X}_i) - Y_i)^2 - \inf_{f \in \mathcal{F}} \sum_{i=1}^n (f(\boldsymbol{X}_i) - Y_i)^2\right] \\
&+ \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \inf_{f \in \mathcal{F}} (f(\boldsymbol{X}_i) - Y_i)^2 - \sum_{i=1}^n (\tilde{f}(\boldsymbol{X}_i) - Y_i)^2\right] \\
&+ \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^n \epsilon_i(\hat{f}(\boldsymbol{X}_i) - \tilde{f}(\boldsymbol{X}_i))\right] \\
=& \Delta_n(\hat{f}, f_0, \mathcal{F}) - \Delta_n(\tilde{f}, f_0, \mathcal{F}) + \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[\epsilon_i \hat{f}(\boldsymbol{X}_i) - \epsilon_i \tilde{f}(\boldsymbol{X}_i)\right].
\end{aligned} \tag{10}$$

Next, note that by the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left|\frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[\epsilon_i \hat{f}(\boldsymbol{X}_i)\right]\right| &\leq \frac{2}{n}\sum_{i=1}^n \sqrt{\mathbb{E}\left[\epsilon_i^2\right]\mathbb{E}\left[\hat{f}(\boldsymbol{X}_i)^2\right]} \\
&\leq \frac{2}{n}\sum_{i=1}^n \sqrt{1 \cdot F^2} \\
&= 2F,
\end{aligned}$$

where the second inequality follows from the fact that ϵ_i is standard normal and $||\hat{f}||_\infty \leq F$. Note that the same bound clearly holds if $\hat{f}$ is replaced by $\tilde{f}$. Hence, using this together with $\Delta_n \geq 0$, $\hat{R}_n(\hat{f}, f_0) \leq 4F^2$ and (10), we can now obtain

$$\begin{aligned}\Delta_n &= \hat{R}_n(\hat{f}, f_0) - \hat{R}_n(\tilde{f}, f_0) + \Delta_n(\tilde{f}, f_0, \mathcal{F}) - \frac{2}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[\epsilon_i \hat{f}(\boldsymbol{X}_i) - \epsilon_i \tilde{f}(\boldsymbol{X}_i)\right]\right) \\ &\leq 4F^2 + \kappa + 2 \cdot 2F \\ &\leq 8F^2 + \kappa,\end{aligned}$$

where the last inequality follows from the fact that $F \geq 1$, so that $F^2 \geq F$. Since this bound on Δ_n holds for any $\kappa > 0$, we have that $\Delta_n \leq 8F^2$. Now if $\log \mathcal{N}_n \geq n$, we obtain that

$$(1-\epsilon)^2\Delta_n - F^2\frac{18\log\mathcal{N}_n + 76}{n\epsilon} - 38\delta F \quad \leq \quad 8F^2 - F^2\frac{18n}{n} \quad \leq \quad 0 \quad \leq \quad R(\hat{f}, f_0).$$

Hence, the lower bound in Lemma 4 holds. The same argument can be used for the lower bound in Lemma 4′. This proves Lemmas 4 and 4′ in the case that $\log \mathcal{N}_n \geq n$.

Now suppose that $\log \mathcal{N}_n \leq n$. The proof in this case is much more involved and will involve four intermediate steps. The general idea of the four steps given by Schmidt-Hieber (2020) is as follows:

(I) We relate $R(\hat{f}, f_0)$ to $\hat{R}_n(\hat{f}, f_0)$ by proving the following bounds:

$$\begin{aligned}(1-\epsilon)\hat{R}_n(\hat{f}, f_0) - F^2\frac{15\log\mathcal{N}_n + 75}{n\epsilon} - 26\delta F \quad &\leq \quad R(\hat{f}, f_0) \\ \leq \quad (1+\epsilon)\left(\hat{R}_n(\hat{f}, f_0) + (1+\epsilon)F^2\frac{12\log\mathcal{N}_n + 70}{n\epsilon} + 26\delta F\right).&\end{aligned}$$

(II) We prove the following bound for any estimator $\tilde{f}$ taking values in $\mathcal{F}$:

$$\left|\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\epsilon_i \tilde{f}(\boldsymbol{X}_i)\right]\right| \leq 2\sqrt{\frac{\hat{R}_n(\tilde{f}, f_0)(3\log\mathcal{N}_n + 1)}{n}} + 6\delta.$$

(III) Using (II), we relate $\hat{R}_n(\hat{f}, f_0)$ to $\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2]$ by proving the following upper bound:

$$\hat{R}_n(\hat{f}, f_0) \leq (1+\epsilon)\left[\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + 6\delta + F^2\frac{6\log\mathcal{N}_n + 2}{n\epsilon} + \Delta_n\right].$$

(IV) We prove the following lower bound for $\hat{R}_n(\hat{f}, f_0)$:

$$\hat{R}_n(\hat{f}, f_0) \geq (1-\epsilon)\left[\Delta_n - \frac{3\log \mathcal{N}_n + 1}{n\epsilon} - 12\delta\right].$$

Schmidt-Hieber (2020) then combines (I) and (IV) to obtain the lower bound in Lemma 4 and (I) and (III) to obtain the upper bound. However, we were unable to prove (I). We suspect that Schmidt-Hieber (2020) made an error in his application of Bernstein's inequality. Instead, we will prove

$$\begin{aligned}(1-\epsilon)\hat{R}_n(\hat{f}, f_0) - &\frac{F^4}{\epsilon n}(15\log\mathcal{N}_n + 43) - 25\delta F^2 \leq R(\hat{f}, f_0)\\ &\leq (1+\epsilon)\left(\hat{R}_n(\hat{f}, f_0) + \frac{(1+\epsilon)F^4}{\epsilon n}(12\log\mathcal{N}_n + 38) + 25\delta F^2\right),\end{aligned}$$

which we will denote (I'). This is the difference that leads to us replacing (8) in Lemma 4 by (9). We now prove the four steps.

(I'): Given a minimal covering of $\mathcal{F}$ by closed balls with radius δ, denote the centers by $(f_j)_{j=1}^{\mathcal{N}_n}$. We can assume without loss of generality that $||f_j||_\infty \leq F$ for all j. Denote by j^* a random index in $[\mathcal{N}_n]$ such that $||f_{j^*} - \hat{f}||_\infty \leq \delta$. Such a j^* exists by definition of the f_j. Let $(\boldsymbol{X}'_i)_{i=1}^n$ be random variables independent of and with the same joint distribution as $(\boldsymbol{X}_i)_{i=1}^n$. For each $j \in [\mathcal{N}_n]$, define a function $g_j(\boldsymbol{x}, \boldsymbol{y}) := (f_j(\boldsymbol{y}) - f_0(\boldsymbol{y}))^2 - (f_j(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2$ for $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$. Note that g_{j^*} is a random function. Then since the $\boldsymbol{X}'_i$ all have the same distribution as $\boldsymbol{X}$, we have

$$\begin{aligned}|R(\hat{f}, f_0) - \hat{R}_n(\hat{f}, f_0)| =& \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(\hat{f}(\boldsymbol{X}'_i) - f_0(\boldsymbol{X}'_i))^2 - \frac{1}{n}\sum_{i=1}^n(\hat{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))^2\right]\right|\\ =& \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(f_{j^*}(\boldsymbol{X}'_i) - f_0(\boldsymbol{X}'_i))^2 - \frac{1}{n}\sum_{i=1}^n(f_{j^*}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))^2\right.\right.\\ &+ \frac{1}{n}\sum_{i=1}^n(\hat{f}(\boldsymbol{X}'_i) + f_{j^*}(\boldsymbol{X}'_i) - 2f_0(\boldsymbol{X}'_i))(\hat{f}(\boldsymbol{X}'_i) - f_{j^*}(\boldsymbol{X}'_i)) \qquad (11)\\ &\left.\left. - \frac{1}{n}\sum_{i=1}^n(\hat{f}(\boldsymbol{X}_i) + f_{j^*}(\boldsymbol{X}_i) - 2f_0(\boldsymbol{X}_i))(\hat{f}(\boldsymbol{X}_i) - f_{j^*}(\boldsymbol{X}_i))\right]\right|\\ \leq& \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n g_{j^*}(\boldsymbol{X}_i, \boldsymbol{X}'_i)\right| + 8\delta F.\end{aligned}$$

The inequality here follows from the definition of g_j, Jensen's inequality, the triangle inequality and the following bound for any $\boldsymbol{x} \in [0,1]^d$:

$$\begin{aligned}\left|(\hat{f}(\boldsymbol{x}) + f_{j^*}(\boldsymbol{x}) - 2f_0(\boldsymbol{x}))(\hat{f}(\boldsymbol{x}) - f_{j^*}(\boldsymbol{x}))\right| &= \left|\hat{f}(\boldsymbol{x}) + f_{j^*}(\boldsymbol{x}) - 2f_0(\boldsymbol{x})\right| \cdot \left|\hat{f}(\boldsymbol{x}) - f_{j^*}(\boldsymbol{x})\right| \\ &\leq 4F \cdot \delta,\end{aligned}$$

where we use the triangle inequality and that $||f_0||_\infty, ||f_j||_\infty \leq F$ for all j and $||f_{j^*} - \hat{f}||_\infty \leq \delta$ by definition of j^*.

Remark A.1. *Schmidt-Hieber (2020) bounds*

$$|R(\hat{f}, f_0) - \hat{R}_n(\hat{f}, f_0)| \leq \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} g_{j^*}(\boldsymbol{X}_i, \boldsymbol{X}_i')\right| + 9\delta F.$$

That is, the term $8\delta F$ in our bound is replaced by $9\delta F$. The inequality clearly still holds, but is weaker.

For any $j \in [\mathcal{N}_n]$, define

$$r_j := \sqrt{n^{-1}\log\mathcal{N}_n} \vee \mathbb{E}^{1/2}\left[(f_j(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right].$$

Let $r^* := r_{j^*}$ be the element of $(r_j)_{j=1}^{\mathcal{N}_n}$ with index j^*. We can then write r^* as

$$r^* = \sqrt{n^{-1}\log\mathcal{N}_n} \vee \mathbb{E}^{1/2}\left[(f_{j^*}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right].$$

Note that $\mathbb{E}^{1/2}\left[(f_{j^*}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right]$ can be viewed as a conditional L^2-distance, which therefore satisfies a conditional triangle inequality:

$$\begin{aligned}&\mathbb{E}^{1/2}[(f_{j^*}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n] \\ &\qquad \leq \mathbb{E}^{1/2}\left[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right] + \mathbb{E}^{1/2}\left[(f_{j^*}(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right] \\ &\qquad \leq \mathbb{E}^{1/2}\left[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right] + \delta,\end{aligned}$$

where the final inequality follows from $||f_{j^*} - f_0||_\infty \leq \delta$. Combining this inequality with the trivial bound $a \vee b \leq a + b$ for any $a, b \geq 0$, we obtain

$$r^* \leq \sqrt{n^{-1}\log \mathcal{N}_n} + \mathbb{E}^{1/2}\left[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right] + \delta.$$

Now define the random variables

$$U := \mathbb{E}^{1/2}\left[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2 | (\boldsymbol{X}_i, Y_i)_{i=1}^n\right], \quad T := \max_{j \in [\mathcal{N}_n]} \frac{|\sum_{i=1}^n g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')|}{r_j F}.$$

Then by the law of iterated expectations, $\mathbb{E}[U^2] = R(\hat{f}, f_0)$. Hence, using the bound on r^*, we obtain

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n g_{j^*}(\boldsymbol{X}_i, \boldsymbol{X}_i')\right| =& \mathbb{E}\left[\frac{1}{n}\frac{|\sum_{i=1}^n g_{j^*}(\boldsymbol{X}_i, \boldsymbol{X}_i')|}{r^* F} r^* F\right] \\
\leq& \mathbb{E}\left[\frac{1}{n} T r^* F\right] \\
\leq& \frac{F}{n}\mathbb{E}\left[T\left(\sqrt{n^{-1}\log \mathcal{N}_n} + U + \delta\right)\right] \\
=& \frac{F}{n}\left(\sqrt{n^{-1}\log \mathcal{N}_n} + \delta\right)\mathbb{E}[T] + \frac{F}{n}\mathbb{E}[TU] \\
\leq& \frac{F}{n}\left(\sqrt{n^{-1}\log \mathcal{N}_n} + \delta\right)\mathbb{E}[T] + \frac{F}{n}\sqrt{\mathbb{E}[T^2]\,\mathbb{E}[U^2]} \\
=& \frac{F}{n}\left(\sqrt{n^{-1}\log \mathcal{N}_n} + \delta\right)\mathbb{E}[T] + \frac{F}{n}\sqrt{\mathbb{E}[T^2]\,R(\hat{f}, f_0)},
\end{aligned}$$

where the final inequality follows from the Cauchy-Schwarz inequality. Combining this with (11) then gives

$$|R(\hat{f}, f_0) - \hat{R}_n(\hat{f}, f_0)| \leq \frac{F}{n}\left(\sqrt{n^{-1}\log \mathcal{N}_n} + \delta\right)\mathbb{E}[T] + \frac{F}{n}\sqrt{\mathbb{E}[T^2]\,R(\hat{f}, f_0)} + 8\delta F. \quad (12)$$

Now note that for any fixed $j \in [\mathcal{N}_n]$, we have that $\mathbb{E}[g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')] = 0$, since $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$ have the same distribution. Furthermore, since $||f_j - f_0||_\infty \leq 2F$ for all j, we have that $|g_j(\boldsymbol{x}, \boldsymbol{y})| \leq 4F^2$ for all j and for all $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$. Hence, $\mathbb{E}[g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')/(r_j F)] = 0$ and $|g_j(\boldsymbol{x}, \boldsymbol{y})/(r_j F)| \leq 4F/r_j$. Also note that $\boldsymbol{X}_i$ and $\boldsymbol{X}_i'$ being independent and both having the same distribution as $\boldsymbol{X}$ implies that

$$\begin{aligned}\operatorname{Var}(g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')) &= 2\operatorname{Var}((f_j(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2) \\ &\leq 2\mathbb{E}\left[(f_j(\boldsymbol{X}) - f_0(\boldsymbol{X}))^4\right] \\ &\leq 2\mathbb{E}\left[(2F)^2(f_j(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right] \\ &\leq 8F^2 r_j^2,\end{aligned}$$

where we again use that $||f_j - f_0||_\infty \leq 2F$ and we use the definition of r_j. Hence, $\operatorname{Var}(g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')/(r_jF)) \leq 8$. We now use the union bound and apply Bernstein's inequality to $(g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')/(r_jF))_{i=1}^n$ for each $j \in [\mathcal{N}_n]$. We then obtain for any $t \geq 0$:

$$\begin{aligned}\mathbb{P}(T \geq t) &= \mathbb{P}\left(\bigcup_{j=1}^{\mathcal{N}_n}\left\{\left|\sum_{i=1}^n \frac{g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')}{r_jF}\right| \geq t\right\}\right) \\ &\leq 1 \wedge \sum_{j=1}^{\mathcal{N}_n}\mathbb{P}\left(\left|\sum_{i=1}^n \frac{g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')}{r_jF}\right| \geq t\right) \\ &\leq 1 \wedge \sum_{j=1}^{\mathcal{N}_n} 2\exp\left(-\frac{t^2}{2\cdot(4F/r_j)\cdot t/3 + 2\sum_{i=1}^n 8}\right) \\ &\leq 1 \wedge 2\mathcal{N}_n \max_{j\in[\mathcal{N}_n]} \exp\left(-\frac{t^2}{8Ft/(3r_j) + 16n}\right).\end{aligned}$$

Remark A.2. *Schmidt-Hieber (2020) writes* $8t/(3r_j) + 16n$ *as the denominator in the final bound, where the bound we obtained has a factor* F *in the first term. We do not see how Schmidt-Hieber (2020) obtains his bound through Bernstein's inequality. We now demonstrate how continuing the proof of Schmidt-Hieber (2020) using our bound rather than his will lead to problems. After demonstrating this, we will show how a different proof can be used to obtain (I').*

Since $r_j \geq \sqrt{n^{-1}\log \mathcal{N}_n}$ by definition, we have for all $t \geq 6\sqrt{n \log \mathcal{N}_n}$ that

$$\begin{aligned}
\mathbb{P}(T \geq t) &\leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t}{8F/(3\sqrt{n^{-1}\log \mathcal{N}_n}) + 16n/t}\right) \\
&\leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t}{8F/(3\sqrt{n^{-1}\log \mathcal{N}_n}) + 16n/(6\sqrt{n\log \mathcal{N}_n})}\right) \\
&\leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t\sqrt{\log \mathcal{N}_n}}{\frac{8}{3}F\sqrt{n} + \frac{8}{3}\sqrt{n}}\right) \\
&= 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{3t\sqrt{\log \mathcal{N}_n}}{8(F+1)\sqrt{n}}\right).
\end{aligned}$$

We now bound $\mathbb{E}[T]$:

$$\begin{aligned}
\mathbb{E}[T] &= \int_0^\infty \mathbb{P}(T \geq t)dt \\
&\leq \int_0^{6\sqrt{n\log \mathcal{N}_n}} 1dt + 2\mathcal{N}_n \int_{6\sqrt{n\log \mathcal{N}_n}}^\infty \exp\left(-\frac{3t\sqrt{\log \mathcal{N}_n}}{8(F+1)\sqrt{n}}\right) dt \\
&=6\sqrt{n\log \mathcal{N}_n} + \mathcal{N}_n \frac{16(F+1)\sqrt{n}}{3\sqrt{\log \mathcal{N}_n}} \exp\left(-\frac{3(6\sqrt{n\log \mathcal{N}_n})\sqrt{\log \mathcal{N}_n}}{8(F+1)\sqrt{n}}\right) \\
&=6\sqrt{n\log \mathcal{N}_n} + \mathcal{N}_n \frac{16(F+1)\sqrt{n}}{3\sqrt{\log \mathcal{N}_n}} \exp\left(-\frac{9}{4(F+1)}\log \mathcal{N}_n\right).
\end{aligned}$$

Remark A.3. *Note that the bound obtained by Schmidt-Hieber (2020) has a factor 2 where we have a factor $F+1$. This factor $F+1$ now leads to problems, because Schmidt-Hieber (2020) can remove the factor $\mathcal{N}_n$ by noting that*

$$\mathcal{N}_n \exp\left(-\frac{9}{8}\log \mathcal{N}_n\right) \leq 1.$$

This conclusion does not necessarily hold for our bound if $F > 5/4$. Hence, we will now take a slightly different approach. This will lead to (I') instead of (I).

Note that (12) can also be written as

$$|R(\hat{f}, f_0) - \hat{R}_n(\hat{f}, f_0)| \leq \frac{F^2}{n}\left(\sqrt{n^{-1}\log \mathcal{N}_n} + \delta\right)\mathbb{E}\left[\frac{T}{F}\right] + \frac{F^2}{n}\sqrt{\mathbb{E}\left[\frac{T^2}{F^2}\right]R(\hat{f}, f_0)} + 8\delta F. \tag{13}$$

We now apply Bernstein's inequality to T/F instead of T itself. Note that $|g_j(\boldsymbol{x}, \boldsymbol{y})|/(F^2 r_j) \leq$

$4/r_j$ and $\mathrm{Var}(g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')/(F^2 r_j)) \leq 8/F^2$. Hence,

$$\begin{aligned}
\mathbb{P}(T/F \geq t) &= \mathbb{P}\left(\bigcup_{j=1}^{\mathcal{N}_n} \left\{\left|\sum_{i=1}^{n} \frac{g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')}{r_j F^2}\right| \geq t\right\}\right) \\
&\leq 1 \wedge \sum_{j=1}^{\mathcal{N}_n} \mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')}{r_j F^2}\right| \geq t\right) \\
&\leq 1 \wedge \sum_{j=1}^{\mathcal{N}_n} 2\exp\left(-\frac{t^2}{2 \cdot (4/r_j) \cdot t/3 + 2\sum_{i=1}^{n} 8/F^2}\right) \\
&\leq 1 \wedge 2\mathcal{N}_n \max_{j \in [\mathcal{N}_n]} \exp\left(-\frac{t^2}{8t/(3r_j) + 16n/F^2}\right).
\end{aligned}$$

Then for $t \geq 6\sqrt{n \log \mathcal{N}_n}$, we obtain in the same way as before:

$$\begin{aligned}
\mathbb{P}(T/F \geq t) &\leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t}{8/(3\sqrt{n^{-1}\log \mathcal{N}_n}) + 16nF^{-2}/t}\right) \\
&\leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t}{8/(3\sqrt{n^{-1}\log \mathcal{N}_n}) + 16nF^{-2}/(6\sqrt{n \log \mathcal{N}_n})}\right) \\
&\leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t\sqrt{\log \mathcal{N}_n}}{\frac{8}{3}\sqrt{n} + \frac{8}{3}F^{-2}\sqrt{n}}\right) \\
&= 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{3t\sqrt{\log \mathcal{N}_n}}{8(1+F^{-2})\sqrt{n}}\right).
\end{aligned}$$

Using that $F^{-2} \leq 1$, we can now bound $\mathbb{E}[T/F]$ in the same way we bounded $\mathbb{E}[T]$ before:

$$\begin{aligned}
\mathbb{E}[T/F] &= \int_0^\infty \mathbb{P}(T/F \geq t) dt \\
&\leq \int_0^{6\sqrt{n \log \mathcal{N}_n}} 1 dt + 2\mathcal{N}_n \int_{6\sqrt{n \log \mathcal{N}_n}}^\infty \exp\left(-\frac{3t\sqrt{\log \mathcal{N}_n}}{8(1+F^{-2})\sqrt{n}}\right) dt \\
&= 6\sqrt{n \log \mathcal{N}_n} + \mathcal{N}_n \frac{16(1+F^{-2})\sqrt{n}}{3\sqrt{\log \mathcal{N}_n}} \exp\left(-\frac{3(6\sqrt{n \log \mathcal{N}_n})\sqrt{\log \mathcal{N}_n}}{8(1+F^{-2})\sqrt{n}}\right) \\
&= 6\sqrt{n \log \mathcal{N}_n} + \mathcal{N}_n \frac{16(1+F^{-2})\sqrt{n}}{3\sqrt{\log \mathcal{N}_n}} \exp\left(-\frac{9}{4(1+F^{-2})} \log \mathcal{N}_n\right) \\
&\leq 6\sqrt{n \log \mathcal{N}_n} + \mathcal{N}_n \frac{16(1+F^{-2})\sqrt{n}}{3\sqrt{\log \mathcal{N}_n}} \mathcal{N}_n^{-9/8} \\
&\leq 6\sqrt{n \log \mathcal{N}_n} + \frac{32\sqrt{n}}{3\sqrt{\log \mathcal{N}_n}}.
\end{aligned}$$

We can apply a similar approach to bound $\mathbb{E}[(T/F)^2]$:

$$\begin{aligned}\mathbb{E}[(T/F)^2] =& \int_0^\infty \mathbb{P}(T/F \geq \sqrt{t})dt\\ \leq& \int_0^{36n\log\mathcal{N}_n} 1dt + 2\mathcal{N}_n \int_{36n\log\mathcal{N}_n}^\infty \exp\left(-\frac{3\sqrt{t}\sqrt{\log\mathcal{N}_n}}{8(1+F^{-2})\sqrt{n}}\right)dt\\ \leq& 36n\log\mathcal{N}_n + 2^7 n.\end{aligned}$$

The final inequality here follows from applying substitution and integration by parts as follows. Let $a, b, c > 0$ be constants, then

$$\begin{aligned}\int_{b^2}^{c^2} \exp(-\sqrt{t}a)dt &= \int_b^c \exp(-ua)\cdot 2udu\\ &= 2\left[\frac{b}{a}\exp(-ba) - \frac{c}{a}\exp(-ca) + \int_b^c \frac{1}{a}\exp(-ua)du\right]\\ &= \frac{2}{a}\left[b\exp(-ba) - c\exp(-ca) + \frac{1}{a}\exp(-ba) - \frac{1}{a}\exp(-ca)\right]\\ &\xrightarrow{c\to\infty} \frac{2}{a}\left[b\exp(-ba) + \frac{1}{a}\exp(-ba)\right]\\ &= \frac{2}{a^2}(ab+1)\exp(-ba).\end{aligned}$$

Applying this with $b = 6\sqrt{n\log\mathcal{N}_n}$ and $a = \frac{3\sqrt{\log\mathcal{N}_n}}{8(1+F^{-2})\sqrt{n}}$, gives

$$\begin{aligned}&2\mathcal{N}_n \int_{36n\log\mathcal{N}_n}^\infty \exp\left(-\frac{3\sqrt{t}\sqrt{\log\mathcal{N}_n}}{8(1+F^{-2})\sqrt{n}}\right)dt\\ &\qquad= \mathcal{N}_n \frac{128(1+F^{-2})^2 n}{9\log\mathcal{N}_n}\left(\frac{18\log\mathcal{N}_n}{8(1+F^{-2})} + 1\right)\exp\left(-\frac{18\log\mathcal{N}_n}{8(1+F^{-2})}\right)\\ &\qquad\leq \mathcal{N}_n \frac{128(1+F^{-2})^2 n}{9\log\mathcal{N}_n}\left(2\cdot\frac{18\log\mathcal{N}_n}{8(1+F^{-2})} + 1\right)\cdot\mathcal{N}_n^{-9/8}\\ &\qquad\leq 64(1+F^{-2})n \leq 2^7 n\end{aligned}$$

where the first inequality follows from the fact that $\log\mathcal{N}_n \geq \log 3 > 1$ and $F^{-2} \leq 1$. We now combine the moment bounds with (13). Together with $\log\mathcal{N}_n \leq n$ and $F \leq F^2$, this gives

$$
\begin{aligned}
&|R(\hat{f}, f_0) - \hat{R}_n(\hat{f}, f_0)|/F^2 \\
&\qquad \le \frac{1}{n}\left(\sqrt{n^{-1}\log\mathcal{N}_n} + \delta\right)\left(6\sqrt{n\log\mathcal{N}_n} + \frac{32\sqrt{n}}{3\sqrt{\log\mathcal{N}_n}}\right) \\
&\qquad\quad + \frac{1}{n}\sqrt{\left[36n\log\mathcal{N}_n + 2^7 n\right]R(\hat{f}, f_0)} + 8\delta/F \\
&\qquad = \frac{6\log\mathcal{N}_n}{n} + \frac{32}{3n} + \frac{6\sqrt{\log\mathcal{N}_n}\delta}{\sqrt{n}} + \frac{32\delta}{3\sqrt{n\log\mathcal{N}_n}} \\
&\qquad\quad + \frac{1}{n}\sqrt{\left[36n\log\mathcal{N}_n + 2^7 n\right]R(\hat{f}, f_0)} + 8\delta/F \\
&\qquad \le \frac{1}{n}\left(6\log\mathcal{N}_n + 11\right) + 6\delta + 11\delta \\
&\qquad\quad + \frac{1}{n}\sqrt{\left[36n\log\mathcal{N}_n + 2^7 n\right]R(\hat{f}, f_0)} + 8\delta \\
&\qquad = \frac{1}{n}\left(6\log\mathcal{N}_n + 11\right) + \frac{1}{n}\sqrt{\left[36n\log\mathcal{N}_n + 2^7 n\right]R(\hat{f}, f_0)} + 25\delta,
\end{aligned}
\tag{14}
$$

Now let $a, b, d \ge 0, c > 0$ be constants such that $|a - b| \le 2\sqrt{ac} + d$. We claim that for all $\epsilon \in (0, 1]$,

$$
(1-\epsilon)b - d - \frac{c^2}{\epsilon} \le a \le (1+\epsilon)(b+d) + \frac{(1+\epsilon)^2}{\epsilon}c^2. \tag{15}
$$

The upper bound is positive and hence trivially holds if $a = 0$. If $a = 0$, the condition $|a - b| \le d$ implies that the lower bound is negative and hence also trivially holds. Now suppose $a > 0$. To prove the upper bound, note that $|a - b| \le 2\sqrt{ac} + d$ implies $a \le b + 2\sqrt{ac} + d$. It is a straightforward calculus exercise to show that $x + x^{-1} \ge 2$ for all $x > 0$. Choosing $x = \frac{\epsilon\sqrt{a}}{(1+\epsilon)c}$ then gives

$$
2 \quad \le \quad \frac{\epsilon\sqrt{a}}{(1+\epsilon)c} + \frac{(1+\epsilon)c}{\epsilon\sqrt{a}} \quad = \quad (\sqrt{a}c)^{-1}\left(\frac{\epsilon}{1+\epsilon}a + \frac{1+\epsilon}{\epsilon}c^2\right).
$$

Combining this with the bound $a \le b + 2\sqrt{a}c + d$ then gives

$$
a \le b + \frac{\epsilon}{1+\epsilon}a + \frac{1+\epsilon}{\epsilon}c^2 + d.
$$

Solving for a yields the upper bound in the claim. For the lower bound, note that $b \le$

$a+2\sqrt{ac}+d$ and the bound on $2\sqrt{ac}$ then implies that

$$\begin{aligned} b &\leq a + \frac{\epsilon}{1+\epsilon}a + \frac{1+\epsilon}{\epsilon}c^2 + d \\ &= \frac{1+2\epsilon}{1+\epsilon}a + \frac{1+\epsilon}{\epsilon}c^2 + d. \end{aligned}$$

It is a simple calculus exercise to show that

$$1 \geq \frac{1+x}{1+2x} \geq 1-x \quad \text{and} \quad \frac{(1+x)^2}{x(1+2x)} \geq x^{-1}, \quad \forall x \in (0,1].$$

Rewriting the bound on b into a bound on a then gives

$$\begin{aligned} a &\geq \frac{1+\epsilon}{1+2\epsilon}b - \frac{(1+\epsilon)^2}{\epsilon(1+2\epsilon)}c^2 - 1 + \epsilon 1 + 2\epsilon d \\ &\geq (1-\epsilon)b - \frac{1}{\epsilon}c^2 - d, \end{aligned}$$

as desired. This proves the claim

We now apply this claim to

$$a = R(\hat{f}, f_0),\ b = \hat{R}_n(\hat{f}, f_0),\ c = \frac{F^2}{n}\sqrt{9n\log\mathcal{N}_n + 32n},\ d = \frac{F^2}{n}\left(6\log\mathcal{N}_n + 11\right) + 25\delta F^2.$$

Then $|a-b| \leq 2\sqrt{ac}+d$ by (14). Using $(1+\epsilon)/\epsilon \geq 2$, we obtain for the upper bound:

$$\begin{aligned} &R(\hat{f}, f_0) \\ &\quad \leq (1+\epsilon)\left(\hat{R}_n(\hat{f}, f_0) + \frac{F^2}{n}\left(6\log\mathcal{N}_n + 11\right) + 25\delta F^2\right) + \frac{(1+\epsilon)^2F^4}{\epsilon n^2}\left(9n\log\mathcal{N}_n + 32n\right) \\ &\quad \leq (1+\epsilon)\left(\hat{R}_n(\hat{f}, f_0) + \frac{(1+\epsilon)F^4}{\epsilon n}\left(3\log\mathcal{N}_n + 6\right) + 25\delta F^2\right) + \frac{(1+\epsilon)^2F^4}{\epsilon n}\left(9\log\mathcal{N}_n + 32\right) \\ &\quad = (1+\epsilon)\left(\hat{R}_n(\hat{f}, f_0) + \frac{(1+\epsilon)F^4}{\epsilon n}\left(12\log\mathcal{N}_n + 38\right) + 25\delta F^2\right). \end{aligned}$$

Similarly, for the lower bound:

$$
\begin{aligned}
&R(\hat{f}, f_0)\\
&\geq (1-\epsilon)\hat{R}_n(\hat{f}, f_0) - \frac{F^2}{n}(6\log\mathcal{N}_n + 11) - 25\delta F^2 - \frac{F^4}{\epsilon n^2}(9n\log\mathcal{N}_n + 32n)\\
&\geq (1-\epsilon)\hat{R}_n(\hat{f}, f_0) - \frac{F^4}{n}(6\log\mathcal{N}_n + 11) - 25\delta F^2 - \frac{F^4}{\epsilon n}(9\log\mathcal{N}_n + 32)\\
&= (1-\epsilon)\hat{R}_n(\hat{f}, f_0) - \frac{F^4}{\epsilon n}(15\log\mathcal{N}_n + 43) - 25\delta F^2.
\end{aligned}
$$

Hence, instead of the initial goal of (I), we have shown the following inequalities:

$$
\begin{aligned}
(1-\epsilon)\hat{R}_n(\hat{f}, f_0) - &\frac{F^4}{\epsilon n}(15\log\mathcal{N}_n + 43) - 25\delta F^2 \leq R(\hat{f}, f_0)\\
&\leq (1+\epsilon)\left(\hat{R}_n(\hat{f}, f_0) + \frac{(1+\epsilon)F^4}{\epsilon n}(12\log\mathcal{N}_n + 38) + 25\delta F^2\right),
\end{aligned}
$$

which we will denote (I'). Our (I') is very similar to (I) of Schmidt-Hieber (2020). The only differences are that F in (I) is replaced by F^2 in (I') and some constants in (I') are smaller than in (I).

(II): Let $\tilde{f}$ be an estimator taking values in $\mathcal{F}$. Let j' be a random index in $[\mathcal{N}_n]$ such that $||f_{j'} - \tilde{f}||_\infty \leq \delta$. Then we have

$$
\mathbb{E}\left[\sum_{i=1}^n |\epsilon_i| \cdot |\tilde{f}(\boldsymbol{X}_i) - f_{j'}(\boldsymbol{X}_i)|\right] \quad \leq \quad \delta\sum_{i=1}^n \mathbb{E}|\epsilon_i| \quad \leq \quad \delta n,
$$

where the final inequality follows from ϵ_i being standard normal and hence having first absolute moment smaller than 1.

For $j \in \mathcal{N}_n$, define

$$
\xi_j := \begin{cases} \frac{\sum_{i=1}^n \epsilon_i(f_j(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))}{\sqrt{n}||f_j - f_0||_n} & \text{if } ||f_j - f_0||_n \neq 0\\ Z & \text{else,} \end{cases}
$$

where Z is a standard normal random variable independent of $(\boldsymbol{X}_i)_{i=1}^n$. Note that by independence of ϵ_i and X_i,

$$\mathbb{E}[\epsilon_i f_0(\boldsymbol{X}_i)] \quad = \quad \mathbb{E}[\epsilon_i]\mathbb{E}[f_0(\boldsymbol{X}_i)] \quad = \quad 0.$$

Hence, by combining this with the previous inequality, Jensen's inequality and the triangle inequality, we obtain

$$\begin{aligned}
\left|\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\tilde{f}(\boldsymbol{X}_i)\right]\right| &= \frac{2}{n}\left|\mathbb{E}\left[\sum_{i=1}^{n}\epsilon_i(\tilde{f}(\boldsymbol{X}_i) - f_{j'}(\boldsymbol{X}_i) + f_{j'}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))\right]\right| \\
&\leq \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}|\epsilon_i|\cdot|\tilde{f}(\boldsymbol{X}_i) - f_{j'}(\boldsymbol{X}_i)|\right] + \frac{2}{n}\mathbb{E}\left|\sum_{i=1}^{n}\epsilon_i(f_{j'}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))\right| \\
&\leq 2\delta + \frac{2}{n}\mathbb{E}\left|\sum_{i=1}^{n}\epsilon_i(f_{j'}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))\right| \\
&= 2\delta + \frac{2}{\sqrt{n}}\mathbb{E}\left[||f_{j'} - f_0||_n\cdot|\xi_{j'}|\right] \\
&\leq 2\delta + \frac{2}{\sqrt{n}}\mathbb{E}\left[\left(||\tilde{f} - f_0||_n + \delta\right)|\xi_{j'}|\right],
\end{aligned} \tag{16}$$

where the final inequality follows from the triangle inequality for the seminorm $||\cdot||_n$ and the fact that $||f||_n \leq ||f||_\infty$ for any f. Now recall that the $(\epsilon_i)_i$ are iid standard normal independent of $(\boldsymbol{X}_i)_i$. Hence for any $j \in [\mathcal{N}_n]$, conditional on $(\boldsymbol{X}_i)_i$, we have that ξ_j is either the standard normal random variable Z, or a linear combination of iid standard normal random variables. If $||f_j - f_0||_n \neq 0$, this implies that ξ_j has conditional mean zero with conditional variance

$$\sum_{i=1}^{n}\frac{(f_j(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))^2}{n||f_j - f_0||_n^2} = 1.$$

That is, regardless of the value of $||f_j - f_0||_n$, we have

$$\xi_j|(\boldsymbol{X}_i)_{i=1}^n \sim N(0,1).$$

Then by Lemma C.1:

$$\mathbb{E}[\xi_{j'}^2] \quad \leq \quad \mathbb{E}[\max_{j\in[\mathcal{N}_n]}\xi_j^2] \quad = \quad \mathbb{E}[\mathbb{E}[\max_{j\in[\mathcal{N}_n]}\xi_j^2|(\boldsymbol{X}_i)_{i=1}^n]] \quad \leq \quad 3\log\mathcal{N}_n + 1. \tag{17}$$

Hence, using the Cauchy-Schwarz inequality and Jensen's inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(||\tilde{f}-f_0||_n+\delta\right)|\xi_{j'}|\right] &= \mathbb{E}\left[||\tilde{f}-f_0||_n|\xi_{j'}|\right]+\delta\mathbb{E}\left[|\xi_{j'}|\right] \\ &\leq \sqrt{\mathbb{E}\left[||\tilde{f}-f_0||_n^2\right]\mathbb{E}[\xi_{j'}^2]}+\delta\sqrt{\mathbb{E}[\xi_{j'}^2]} \\ &\leq \sqrt{\hat{R}_n(\tilde{f},f_0)(3\log\mathcal{N}_n+1)}+\delta\sqrt{3\log\mathcal{N}_n+1} \\ &= (\sqrt{\hat{R}_n(\tilde{f},f_0)}+\delta)\sqrt{3\log\mathcal{N}_n+1}.\end{aligned} \tag{18}$$

Since $\log\mathcal{N}_n \leq n$, we have that

$$\frac{2}{\sqrt{n}}\delta\sqrt{3\log\mathcal{N}_n+1} \quad\leq\quad \frac{2}{\sqrt{n}}\delta\sqrt{3n+1} \quad\leq\quad \frac{2}{\sqrt{n}}\delta\sqrt{4n} \quad=\quad 4\delta.$$

Now apply this inequality and (18) to (16) to obtain

$$\begin{aligned}\left|\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^n\epsilon_i\tilde{f}(\boldsymbol{X}_i)\right]\right| &\leq 2\delta+\frac{2}{\sqrt{n}}\sqrt{\hat{R}_n(\tilde{f},f_0)}\sqrt{3\log\mathcal{N}_n+1}+\frac{2}{\sqrt{n}}\delta\sqrt{3\log\mathcal{N}_n+1} \\ &\leq 6\delta+\frac{2}{\sqrt{n}}\sqrt{\hat{R}_n(\tilde{f},f_0)(3\log\mathcal{N}_n+1)},\end{aligned}$$

as desired.

(III): Fix some $f\in\mathcal{F}$. By definition of Δ_n, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(Y_i-\hat{f}(\boldsymbol{X}_i))^2\right]=\mathbb{E}\left[\inf_{g\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(Y_i-g(\boldsymbol{X}_i))^2\right]+\Delta_n\leq\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(Y_i-f(\boldsymbol{X}_i))^2\right]+\Delta_n.$$

Recall that $\mathbb{E}[\epsilon_i f_0(\boldsymbol{X}_i)]=0$. The same obviously holds with f_0-f instead of f_0. Note that then

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(Y_i-f(\boldsymbol{X}_i))^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(f_0(\boldsymbol{X}_i)+\epsilon_i-f(\boldsymbol{X}_i))^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(f_0(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right]+\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\epsilon_i^2+2\epsilon_i(f_0(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(f_0(\boldsymbol{X}_i)-f(\boldsymbol{X}_i))^2\right]+1 \\ &= \mathbb{E}[(f(\boldsymbol{X})-f_0(\boldsymbol{X}))^2]+1\end{aligned}$$

where we use standard normality of ϵ_i in the penultimate equality, and the last equality follows from f, f_0 being fixed and $\boldsymbol{X}_i$ having the same distribution as $\boldsymbol{X}$ for each i. These facts, together with (II) and standard normality of ϵ_i, give

$$\begin{aligned}
\hat{R}_n(\hat{f}, f_0) =& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))^2\right] \\
=& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i) - \epsilon_i)^2\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} 2\epsilon_i\hat{f}(\boldsymbol{X}_i) - \epsilon_i^2 - 2\epsilon_i f_0(\boldsymbol{X}_i)\right] \\
=& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(\boldsymbol{X}_i) - Y_i)^2\right] + \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\hat{f}(\boldsymbol{X}_i)\right] - 1 \\
\leq& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2\right] + \Delta_n + \left|\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\hat{f}(\boldsymbol{X}_i)\right]\right| \\
\leq& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2\right] + \Delta_n + \frac{2}{\sqrt{n}}\sqrt{\hat{R}_n(\hat{f}, f_0)(3\log\mathcal{N}_n + 1)} + 6\delta \\
=& \mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + \frac{2}{\sqrt{n}}\sqrt{\hat{R}_n(\hat{f}, f_0)(3\log\mathcal{N}_n + 1)} + 6\delta.
\end{aligned}$$

We can now apply (15) with

$$a = \hat{R}_n(\hat{f}, f_0), \quad b = 0, \quad c = \sqrt{(3\log\mathcal{N}_n + 1)/n}, \quad d = \mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + 6\delta.$$

Using also that $2F^2 \geq 1 + \epsilon$, we then obtain the following upper bound:

$$\begin{aligned}
\hat{R}_n(\hat{f}, f_0) &\leq (1+\epsilon)\big(\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + 6\delta\big) + \frac{(1+\epsilon)^2}{\epsilon n}(3\log\mathcal{N}_n + 1) \\
&\leq (1+\epsilon)\big(\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + 6\delta\big) + \frac{2(1+\epsilon)F^2}{\epsilon n}(3\log\mathcal{N}_n + 1) \\
&\leq (1+\epsilon)\left(\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + 6\delta + \frac{F^2}{\epsilon n}(6\log\mathcal{N}_n + 2)\right).
\end{aligned}$$

Now note that the upper bound holds for all $f \in \mathcal{F}$ and hence we can replace $\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2]$ by $\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2]$. This proves (III).

(IV): Let $\kappa > 0$ and suppose $\tilde{f}$ is an estimator taking values in $\mathcal{F}$ with $n^{-1}\sum_{i=1}^{n}(\tilde{f}(\boldsymbol{X}_i) - Y_i)^2 - \inf_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^{n}(f(\boldsymbol{X}_i) - Y_i)^2 < \kappa$. Then $\Delta_n(\tilde{f}, f_0, \mathcal{F}) < \kappa$ and from (10) and (II) we obtain

$$\begin{aligned}
&\hat{R}_n(\hat{f}, f_0) - \hat{R}_n(\tilde{f}, f_0) \\
&\quad \geq \Delta_n(\hat{f}, f_0, \mathcal{F}) + \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\left[\epsilon_i \hat{f}(\boldsymbol{X}_i)\right] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\left[\epsilon_i \tilde{f}(\boldsymbol{X}_i)\right] - \kappa \\
&\quad \geq \Delta_n(\hat{f}, f_0, \mathcal{F}) - 2\sqrt{\frac{\hat{R}_n(\hat{f}, f_0)(3\log \mathcal{N}_n + 1)}{n}} - 2\sqrt{\frac{\hat{R}_n(\tilde{f}, f_0)(3\log \mathcal{N}_n + 1)}{n}} - 12\delta - \kappa.
\end{aligned}$$

Now note that for any constants $a, c \geq 0$, we have $2ac \leq \frac{\epsilon}{1-\epsilon}a + \frac{1-\epsilon}{\epsilon}c^2$ and, in particular, $2ac \leq a^2 + c^2$. Combining this with $\frac{1-\epsilon}{\epsilon} + 1 = 1/\epsilon$ then gives

$$\begin{aligned}
&2\sqrt{\frac{\hat{R}_n(\hat{f}, f_0)(3\log \mathcal{N}_n + 1)}{n}} + 2\sqrt{\frac{\hat{R}_n(\tilde{f}, f_0)(3\log \mathcal{N}_n + 1)}{n}} \\
&\leq \frac{\epsilon}{1-\epsilon}\hat{R}_n(\hat{f}, f_0) + \frac{1-\epsilon}{\epsilon} \cdot \frac{3\log \mathcal{N}_n + 1}{n} + \hat{R}_n(\tilde{f}, f_0) + \frac{3\log \mathcal{N}_n + 1}{n} \\
&= \frac{\epsilon}{1-\epsilon}\hat{R}_n(\hat{f}, f_0) + \frac{3\log \mathcal{N}_n + 1}{n\epsilon} + \hat{R}_n(\tilde{f}, f_0).
\end{aligned}$$

Hence, we now obtain

$$\begin{aligned}
&\hat{R}_n(\hat{f}, f_0) - \hat{R}_n(\tilde{f}, f_0) \\
&\quad \geq \Delta_n(\hat{f}, f_0, \mathcal{F}) - \frac{\epsilon}{1-\epsilon}\hat{R}_n(\hat{f}, f_0) - \frac{3\log \mathcal{N}_n + 1}{n\epsilon} - \hat{R}_n(\tilde{f}, f_0) - 12\delta - \kappa.
\end{aligned}$$

Note that the $\hat{R}_n(\tilde{f}, f_0)$ terms cancel and solving for $\hat{R}_n(\hat{f}, f_0)$ then gives

$$\hat{R}_n(\hat{f}, f_0) \geq (1-\epsilon)\left(\Delta_n - \frac{3\log \mathcal{N}_n + 1}{n\epsilon} - 12\delta - \kappa\right).$$

Since this holds for all $\kappa > 0$, we have in particular that

$$\hat{R}_n(\hat{f}, f_0) \geq (1-\epsilon)\left(\Delta_n - \frac{3\log \mathcal{N}_n + 1}{n\epsilon} - 12\delta\right),$$

as desired. This proves (IV).

We can now combine (I)-(IV) to prove the lemma. From (I') and (III), we obtain the following upper bound:

$$R(\hat{f}, f_0) \leq (1+\epsilon)^2 \left(\inf_{f \in \mathcal{F}} \mathbb{E}[(\hat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2] + \Delta_n + \frac{F^4}{\epsilon n} (18 \log \mathcal{N}_n + 40) + 31\delta F^2 \right),$$

using that $F^4 \geq F^2 \geq 1$ and $1 + \epsilon > 1$. From (I') and (IV), we obtain the following lower bound:

$$R(\hat{f}, f_0) \geq (1-\epsilon)^2 \Delta_n - \frac{F^4}{\epsilon n} (18 \log \mathcal{N}_n + 44) - 37\delta F^2,$$

using that $F^4 \geq F^2 \geq 1$ and $(1-\epsilon)^2 \leq 1$. Hence, we have now proven (9). The only differences with (8) are that in our result, the F factors in (8) have been replaced by F^2, and that our constants are smaller. If (I) is used instead of (I') in the above, we obtain the original bounds (8) as given in Lemma 4 by Schmidt-Hieber (2020). □

All that is left to obtain Theorem $2'$ from Lemma $4'$ is a bound on $\log \mathcal{N}_n$ for $\mathcal{F} = \mathcal{F}(L, \boldsymbol{p}, s, F)$. This is the point of Lemma 5.

Lemma 5. *Let $(L, \boldsymbol{p})$ be a network architecture and $s \in \mathbb{N}$. Define $V := \prod_{\ell=0}^{L+1} (p_\ell + 1)$. Then for any $\delta > 0$,*

$$\log \mathcal{N}(\delta, \mathcal{F}_0(L, \boldsymbol{p}, s, \infty), || \cdot ||_\infty) \leq (s+1) \log \left(2\delta^{-1} (L+1) V^2 \right).$$

Proof. For any NN $f = W_L \sigma_{\boldsymbol{v}_L} ... W_1 \sigma_{\boldsymbol{v}_1} W_0$ and any $k \in [L]$, define

$$A_k^+ f : [0,1]^{p_0} \to \mathbb{R}^{p_k}, \quad A_k^+ f := \sigma_{\boldsymbol{v}_k} W_{k-1} \sigma_{\boldsymbol{v}_{k-1}} ... W_1 \sigma_{\boldsymbol{v}_1} W_0,$$

and

$$A_k^- f : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_{L+1}}, \quad A_k^- f := W_L \sigma_{\boldsymbol{v}_L} ... W_k \sigma_{\boldsymbol{v}_k} W_{k-1}.$$

Note that these definitions actually depend on the particular representation of f that we have chosen. That is, the choice of W_i and $\sigma_{\boldsymbol{v}_i}$. This representation is usually not unique, but the dependence on the chosen representation is omitted from notation. Let $A_{L+1}^- f := W_L$. Further define both $A_0^+ f$ and $A_{L+2}^- f$ to be the identity. Now suppose $f \in \mathcal{F}_0(L, \boldsymbol{p})$ and let $k \in [L]$. Then the network parameters are bounded in absolute value by one and hence it can easily be shown by induction on k that $||A_k^+ f||_\infty \leq \prod_{\ell=0}^{k-1} (p_\ell + 1)$.

Recall that the composition of an a_1-Lipschitz function with an a_2-Lipschitz function results in an $a_1 a_2$-Lipschitz function. Now on any subset of $\mathbb{R}^m$, consider the metric induced by the norm $|\cdot|_\infty$. Then any $(r \times \ell)$ real matrix W satisfies $|W\boldsymbol{x}|_\infty \leq \ell ||W||_\infty |\boldsymbol{x}|_\infty$ for all $\boldsymbol{x} \in \mathbb{R}^\ell$. From this, it follows that W is $\ell||W||_\infty$-Lipschitz when considered as a linear function. Note further that $\sigma_{\boldsymbol{v}}$ is 1-Lipschitz for any $\boldsymbol{v}$. Since $||W_i||_\infty \leq 1$, this implies that $A_k^- f$ is $\left(\prod_{\ell=k-1}^{L} p_\ell\right)$-Lipschitz.

Now fix some $\epsilon > 0$ and let $f, f^* \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ with representations

$$f = W_L \sigma_{\boldsymbol{v}_L} ... W_1 \sigma_{\boldsymbol{v}_1} W_0, \quad f^* = W_L^* \sigma_{\boldsymbol{v}_L^*} ... W_1^* \sigma_{\boldsymbol{v}_1^*} W_0^*,$$

and suppose $||W_i - W_i^*||_\infty, |\boldsymbol{v}_i - \boldsymbol{v}_i^*|_\infty \leq \epsilon$ for all i. With both $\sigma_{\boldsymbol{v}_{L+1}}$ and $\sigma_{\boldsymbol{v}_{L+1}^*}$ defined as the identity, we can consider the following telescoping sum:

$$f - f^* = \sum_{k=1}^{L+1} \Big(\left(A_{k+1}^- f\right) \sigma_{\boldsymbol{v}_k} W_{k-1} \left(A_{k-1}^+ f^*\right) - \left(A_{k+1}^- f\right) \sigma_{\boldsymbol{v}_k^*} W_{k-1}^* \left(A_{k-1}^+ f^*\right) \Big).$$

Note that for any $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^m$, we have that $|\sigma_{\boldsymbol{v}} \boldsymbol{x} - \sigma_{\boldsymbol{w}} \boldsymbol{x}|_\infty \leq |\boldsymbol{v} - \boldsymbol{w}|_\infty$. Combining the triangle equality and the properties above, we obtain for any $\boldsymbol{x} \in [0,1]^d$:

$$
\begin{aligned}
|f(\boldsymbol{x}) - f^*(\boldsymbol{x})|_\infty \leq & \sum_{k=1}^{L+1} \left| \left(A_{k+1}^- f\right) \sigma_{\boldsymbol{v}_k} W_{k-1} \left(A_{k-1}^+ f^*\right) \boldsymbol{x} - \left(A_{k+1}^- f\right) \sigma_{\boldsymbol{v}_k^*} W_{k-1}^* \left(A_{k-1}^+ f^*\right) \boldsymbol{x} \right|_\infty \\
\leq & \sum_{k=1}^{L+1} \left(\prod_{\ell=k}^{L} p_\ell \right) \left| \sigma_{\boldsymbol{v}_k} W_{k-1} \left(A_{k-1}^+ f^*\right) \boldsymbol{x} - \sigma_{\boldsymbol{v}_k^*} W_{k-1}^* \left(A_{k-1}^+ f^*\right) \boldsymbol{x} \right|_\infty \\
\leq & \sum_{k=1}^{L+1} \left(\prod_{\ell=k}^{L} p_\ell \right) \Bigg(\left| \sigma_{\boldsymbol{v}_k} W_{k-1} \left(A_{k-1}^+ f^*\right) \boldsymbol{x} - \sigma_{\boldsymbol{v}_k^*} W_{k-1} \left(A_{k-1}^+ f^*\right) \boldsymbol{x} \right|_\infty \\
& + \left| \sigma_{\boldsymbol{v}_k^*} W_{k-1} \left(A_{k-1}^+ f^*\right) \boldsymbol{x} - \sigma_{\boldsymbol{v}_k^*} W_{k-1}^* \left(A_{k-1}^+ f^*\right) \boldsymbol{x} \right|_\infty \Bigg) \\
\leq & \sum_{k=1}^{L+1} \left(\prod_{\ell=k}^{L} p_\ell \right) \left(|\boldsymbol{v}_k - \boldsymbol{v}_k^*|_\infty + \left| \left(W_{k-1} - W_{k-1}^*\right) \left(A_{k-1}^+ f^*\right) \boldsymbol{x} \right|_\infty \right) \\
\leq & \sum_{k=1}^{L+1} \left(\prod_{\ell=k}^{L} p_\ell \right) \left(\epsilon + p_{k-1} ||W_{k-1} - W_{k-1}^*||_\infty \left| \left(A_{k-1}^+ f^*\right) \boldsymbol{x} \right|_\infty \right) \\
\leq & \sum_{k=1}^{L+1} \left(\prod_{\ell=k}^{L} p_\ell \right) \left(\epsilon + p_{k-1} \epsilon \left(\prod_{\ell=0}^{k-2} (p_\ell + 1) \right) \right) \\
= & \epsilon \sum_{k=1}^{L+1} \left(\left(\prod_{\ell=k}^{L} p_\ell \right) + \left(\prod_{\ell=k}^{L} p_\ell \right) p_{k-1} \prod_{\ell=0}^{k-2} (p_\ell + 1) \right) \\
\leq & \epsilon \sum_{k=1}^{L+1} \left(\prod_{\ell=k}^{L} p_\ell \right) (p_{k-1} + 1) \prod_{\ell=0}^{k-2} (p_\ell + 1) \\
\leq & \epsilon \sum_{k=1}^{L+1} V = \epsilon (L+1) V.
\end{aligned}
$$

Now note that a network with architecture $(L, \boldsymbol{p})$ has $\sum_{\ell=0}^{L} p_\ell p_{\ell+1}$ weight parameters and $\sum_{\ell=1}^{L} p_\ell$ shift parameters, for a total number of parameters of

$$
\begin{aligned}
\sum_{\ell=0}^{L} p_\ell p_{\ell+1} + \sum_{\ell=1}^{L} p_\ell &= \sum_{\ell=0}^{L} p_\ell p_{\ell+1} + \sum_{\ell=0}^{L-1} p_{\ell+1} \\
&= \sum_{\ell=0}^{L} (p_\ell + 1) p_{\ell+1} - p_{L+1}.
\end{aligned}
$$

Hence, the total number of parameters can be bounded from above by

$$
\sum_{\ell=0}^{L} (p_\ell + 1) p_{\ell+1} \quad \leq \quad \sum_{\ell=0}^{L} 2^{-L} V \quad = \quad (L+1) 2^{-L} V \quad \leq \quad V,
$$

where we use in the first inequality that

$$V \quad = \quad (p_k+1)(p_{k+1}+1) \prod_{\substack{\ell \in [L+1]_0 \\ \ell \notin \{k,k+1\}}} (p_\ell + 1) \quad \geq \quad (p_k+1)p_{k+1}2^L,$$

for any $k \in [L]_0$. This implies that there are less than V^s ways to choose s network parameters that may be non-zero. Next, we discretize the range $[-1,1]$ of the non-zero parameters using a regular grid with step size $\delta/((L+1)V)$. That is, we consider the grid $G := \{-1 + i \cdot \delta/((L+1)V) : i \in [\underline{2(L+1)V/\delta}]\}$. Note that $\delta/((L+1)V) < 2$ and the grid contains $\underline{2(L+1)V/\delta}$ points. Hence, for each choice of s parameters, this leads to at most $(2(L+1)V/\delta)^s$ ways to choose values in G for those s parameters. Now let P denote the total number of network parameters of a network with architecture $(L, \boldsymbol{p})$ and define the following subset of $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$:

$$\mathcal{S} := \{W_L \sigma_{\boldsymbol{v}_L} ... W_1 \sigma_{\boldsymbol{v}_1} W_0 \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty) : \text{ at least } s \wedge P \text{ network parameters are in } G\}.$$

That is, $\mathcal{S}$ consists of all networks in $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ that can be constructed by choosing s network parameters, letting these s parameters take values in G, and setting all other network parameters to zero. In the case that $P < s$, we let all P network parameters take value in G. By the above, $\mathcal{S}$ has a cardinality of at most $V^s \, (2(L+1)V/\delta)^s = (2(L+1)V^2/\delta)^s$.

We now show $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ can be covered by the set of closed $|| \cdot ||_\infty$-balls of radius δ with centers in $\mathcal{S}$. Indeed, let

$$f = W_L \sigma_{\boldsymbol{v}_L} ... W_1 \sigma_{\boldsymbol{v}_1} W_0 \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty).$$

Note that the grid G has the property that for any $r \in [-1,1]$, there exists $r^* \in G$ such that $|r - r^*| \leq \delta/((L+1)V)$. Note that this is in particular also true for $r = 0$. We use this to construct a network $f^* \in \mathcal{S}$ to approximate f. By assumption, (this representation of) f has exactly m non-zero parameters for some $m \leq s \wedge P$. We can approximate these parameter values by values in G such that the approximation error for each parameter is at most $\delta/((L+1)V)$. Furthermore, we can approximate $(s \wedge P) - m$ of the parameters in f that are zero by choosing a parameter value in $r^* \in G$ with $|r^*| \leq \delta/((L+1)V)$. Set all other $P - (s \wedge P)$ parameters of f^* to zero. This leads to at least $s \wedge P$ network parameters that take value in G and at most $s \wedge P \leq s$ non-zero parameters. Hence, we have found an approximation $f^* = W_L^* \sigma_{\boldsymbol{v}_L^*} ... W_1^* \sigma_{\boldsymbol{v}_1^*} W_0^* \in \mathcal{S}$ of f such that $||W_i - W_i^*||_\infty, |\boldsymbol{v}_i - \boldsymbol{v}_i^*|_\infty \leq \delta/((L+1)V)$ for all i. Finally, we have by the bound we proved above that $||f - f^*||_\infty \leq (L+1)V\delta/((L+1)V) = \delta$.

This proves that $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ can be covered by the set of closed $|| \cdot ||_\infty$-balls of radius δ with centers in $\mathcal{S}$. To complete the proof, note that

$$\log\left(\mathcal{N}(\mathcal{F}_0(L, \boldsymbol{p}, s, \infty))\right) \leq \log\left(|\mathcal{S}|\right) \leq \log\left((2(L+1)V^2/\delta)^s\right) = s \log\left(2\delta^{-1}(L+1)V^2\right).$$

Remark A.4. *The bound we have obtained here is slightly sharper than the bound given in Lemma 5 by Schmidt-Hieber (2020), since our bound has a factor s instead of a factor $s+1$. This is because Schmidt-Hieber (2020) does not consider $\mathcal{S}$, as we do, but instead considers a set with cardinality bounded by*

$$\sum_{s^*=0}^{s} (2(L+1)V^2/\delta)^{s^*} \leq (2(L+1)V^2/\delta)^{s+1}.$$

Schmidt-Hieber (2020) does not write down exactly what set he considers, but considering $\mathcal{S}$ instead leads to a sharper bound, as our proof shows. This sharper bound can be used to prove a slightly sharper version of Theorem 2′, but does not lead to any change in Theorem 1′.

□

We are now ready to prove Theorem 2′ by combining Lemmas 4′ and 5.

Proof of Theorem 2′. We will prove Theorem 2′ assuming Lemma 4′ holds. It is easy to modify this proof to see how Schmidt-Hieber (2020) originally obtained Theorem 2 from Lemma 4 instead of Lemma 4′.

Let $\delta := 1/n$ and $\mathcal{F} := \mathcal{F}_0(L, \boldsymbol{p}, s, F)$. We show that (9) holds. If $\mathcal{N}_n \geq 3$, with $\mathcal{N}_n$

defined as in Lemma 4, this follows immediately from Lemma 4$'$. If $s = 1$, we have that $\mathcal{F}$ contains only the zero function $z : \boldsymbol{x} \mapsto 0$, since $L \geq 1$. Then $\Delta_n = 0$ and

$$0 \leq R(\hat{f}, f_0) = \mathbb{E}[f_0(\boldsymbol{X})^2] = \inf_{f \in \mathcal{F}} \mathbb{E}[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2],$$

which together imply that (9) holds.

Now suppose $n \geq 3$ and $s \geq 2$. Then $\delta \leq 1/3$. We show that $\mathcal{N}_n \geq 3$. Note that $s \geq 2$ implies that $\mathcal{F}$ contains any constant function that takes value in $[-1, 1]$. Indeed, for any $c \in [-1, 1]$, set a shift parameter in the last hidden layer to be $-|c|$ and set the corresponding weight between the last hidden layer and the output layer to be $\operatorname{sgn}(c)$. Set all other network parameters to be 0. This network is then the constant function with value c. In particular, $\mathcal{F}$ contains the three constant functions that take value $-1, 0$ and 1. It is easy to see that this requires at least three balls of radius $\delta \leq 1/3$ to cover. Hence, $\mathcal{N}_n \geq 3$ and (9) holds.

Now suppose $n \leq 2$, then it follows from $||f_0||_\infty \leq F$ and $||f||_\infty \leq F$ for all $f \in \mathcal{F}$ that $0 \leq R(\hat{f}, f_0) \leq 4F^2$ and that for any $f \in \mathcal{F}$ and $i \in [n]$:

$$\begin{aligned}\left|(\hat{f}(\boldsymbol{X}_i) - Y_i)^2 - (f(\boldsymbol{X}_i) - Y_i)^2\right| &= \left|\hat{f}(\boldsymbol{X}_i)^2 - f(\boldsymbol{X}_i)^2 - 2Y_i\left(\hat{f}(\boldsymbol{X}_i) - f(\boldsymbol{X}_i)\right)\right| \\ &\leq \left|\hat{f}(\boldsymbol{X}_i)\right|^2 + |f(\boldsymbol{X}_i)|^2 + 2|Y_i|\left(\left|\hat{f}(\boldsymbol{X}_i)\right| + |f(\boldsymbol{X}_i)|\right) \\ &\leq F^2 + F^2 + 2|Y_i|(2F) = 2F^2 + 4F|Y_i|,\end{aligned}$$

where we use the triangle inequality. Combine this with

$$\mathbb{E}|Y_i| \leq \mathbb{E}|f_0(\boldsymbol{X}_i)| + \mathbb{E}|\epsilon_i| \leq F + 1,$$

and $F \leq F^2 \leq F^4$ to obtain that $\Delta_n \leq 6F^2 + 4F \leq 10F^4$. Hence,

$$(1-\epsilon)^2\Delta_n - F^4\frac{18\log\mathcal{N}_n + 44}{n\epsilon} - 37\delta F^2 \leq 10F^4 - F^4\frac{44}{2} \leq 0,$$

and

$$(1+\epsilon)^2\left[\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\boldsymbol{X})-f_0(\boldsymbol{X}))^2]+F^4\frac{18\log\mathcal{N}_n+40}{n\epsilon}+31\delta F^2+\Delta_n\right]\geq F^4\frac{40}{2}\geq 4F^4.$$

These then imply that (9) holds if $n\leq 2$. Together with the case $n\geq 3$ treated above, we have now shown that (9) always holds.

Before we combine this with Lemma 5, observe that there can never be more than s active nodes in a layer, because each active node requires a non-zero weight corresponding to it. Hence,

$$\mathcal{F}(L,\boldsymbol{p},s,F)=\mathcal{F}(L,(p_0,p_1\wedge s,...,p_L\wedge s,1),s,F).$$

and

$$\begin{aligned}\mathcal{N}_n&=\mathcal{N}(\delta,\mathcal{F}_0(L,(p_0,p_1\wedge s,...,p_L\wedge s,1),s,F),||\cdot||_\infty)\\&\leq\mathcal{N}(\delta,\mathcal{F}_0(L,(p_0,p_1\wedge s,...,p_L\wedge s,1),s,\infty),||\cdot||_\infty).\end{aligned}$$

Then by Lemma 5,

$$\log\mathcal{N}_n\leq(s+1)\log(2\delta^{-1}(L+1)W^2),$$

where $W:=\prod_{\ell=0}^{L+1}\left((p_\ell\wedge s)+1\right)$. We can bound W by

$$W\leq(p_0+1)(p_{L+1}+1)\prod_{\ell=1}^{L}(s+1)\leq 2^2p_0p_{L+1}(s+1)^L.$$

Together with $p_{L+1}=1$ and $\delta=1/n$, this then gives for any $k\in\mathbb{N}_0$ that

$$\begin{aligned}\log\mathcal{N}_n+k&\leq(s+1)\log(2^5n(L+1)(s+1)^{2L}p_0^2)+k\\&\leq 2(s+1)\log(2^3n(L+1)(s+1)^Lp_0)+k\\&\leq 2(s+1)\left[\log(n(L+1)(s+1)^Lp_0)+3\log(2)+\frac{k}{2(s+1)\log(2)}\log(2)\right]\\&\leq 2(s+1)\left[\left(4+\frac{k}{2(s+1)\log(2)}\right)\log(n(L+1)(s+1)^Lp_0)\right]\\&\leq 2\left(4+\frac{k}{2\log(2)}\right)(s+1)\log(n(L+1)(s+1)^Lp_0),\end{aligned}$$

where the penultimate inequality follows from $n(L+1)(s+1)^Lp_0\geq 2$. We now apply this inequality to (9), with $\delta=1/n$ and $F\geq 1$. Let C denote a generic positive constant that depends only on ϵ and might change value from line to line. For the lower bound, we now obtain

$$\begin{aligned} F^4\frac{18\log\mathcal{N}_n+44}{n\epsilon}+37\delta F^2 &\le F^4\frac{18\log\mathcal{N}_n+44}{n\epsilon}+37\frac{F^2}{n\epsilon} \\ &\le CF^4\frac{(s+1)\log(n(L+1)(s+1)^Lp_0)}{n}, \end{aligned}$$

as desired. Similarly, we obtain for the upper bound that

$$(1+\epsilon)^2\left[F^4\frac{18\log\mathcal{N}_n+40}{n\epsilon}+31\delta F^2\right]\le CF^4\frac{(s+1)\log(n(L+1)(s+1)^Lp_0)}{n}.$$

Note that this upper bound can be upper bounded by $\tau_{\epsilon,n}$ for a large enough choice of C_ϵ, since $s, L\ge 1$ and hence $(s+1)^L\ge 2^L\ge L+1$.

Finally, observe that

$$\inf_{f\in\mathcal{F}}\mathbb{E}[(f(\boldsymbol{X})-f_0(\boldsymbol{X}))^2]\le\inf_{f\in\mathcal{F}}||f-f_0||_\infty^2.$$

This completes the proof.

□

We will combine Theorem 2′, Theorem 5 and Lemma 3′ to prove Theorem 1′. Before we can state Lemmas 3 and 3, some context is necessary. Let $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}$ be as in (2) and let $(K_i)_{i=0}^q$ be constants with $K_i\ge 1$. Let $d_0=d$, $d_{q+1}=1$ and let $f=g_q\circ...\circ g_0$ be as in (2), except that we require $g_{ij}\in\mathcal{C}_{t_i}^{\beta_i}([a_i,b_i]^{t_i},K_i)$ and $|a_i|,|b_i|\le K_i$ for all i,j, instead of $g_{ij}\in\mathcal{C}_{t_i}^{\beta_i}([a_i,b_i]^{t_i},K)$ and $|a_i|,|b_i|\le K$. Note that this implies for all i,j that $|g_{ij}|\le K_i$ everywhere. For a vector $\boldsymbol{x}=(x_i)_{i=1}^m\in\mathbb{R}^m$, write $\boldsymbol{x}+c:=(x_i+c)_{i=1}^m$ for any $c\in\mathbb{R}$. Now define

$$h_0:=\frac{g_0}{2K_0}+\frac{1}{2}\quad\text{ and }\quad h_q:=(\boldsymbol{x}\mapsto g_q(2K_{q-1}\boldsymbol{x}-K_{q-1})),$$

as functions $[0,1]^{d_0}\to[0,1]^{d_1}$ and $[0,1]^{d_q}\to\mathbb{R}$ respectively. For $i\in[q-1]$, define

$$h_i:[0,1]^{d_i}\to[0,1]^{d_{i+1}},\quad \boldsymbol{x}\mapsto\frac{g_i(2K_{i-1}\boldsymbol{x}-K_{i-1})}{2K_i}+\frac{1}{2}.$$

It is easy to see by induction that $g_q\circ...\circ g_0=h_q\circ...\circ h_0$. That is, $f=h_q\circ...\circ h_0$.

Now suppose $g \in \mathcal{C}_r^\beta(D, K)$ for some $\beta > 0, r \in \mathbb{R}, D \subset \mathbb{R}^r$ and $K > 0$. Then for any constants $c_1, c_2 \in \mathbb{R}$, we have that $c_1 g + c_2$ still has Hölder smoothness index β, but the Hölder norm changes: $c_1 g + c_2 \in \mathcal{C}_r^\beta(D, |c_1|K + |c_2|)$. Similarly, if $c_1 \geq 1$, we have that $\tilde{g} : \boldsymbol{x} \mapsto g(c_1 \boldsymbol{x} + c_2)$ still has the same Hölder smoothness index β, but the domain and Hölder norm change: $\tilde{g} \in \mathcal{C}_r^\beta((D - c_2)/c_1, c_1^\beta K)$. These results can easily be derived from the definition of Hölder continuity and the Hölder norm.

In the case of the h_i we just defined, these results, together with $K_i \geq 1$ for all i, imply that $h_{0j} \in \mathcal{C}_{t_0}^{\beta_0}([0,1]^{t_0}, 1)$ for all j, $h_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([0,1]^{t_i}, (2K_{i-1})^{\beta_i})$ for all $i \in [q-1]$ and j, and $h_{qj} \in \mathcal{C}_{t_q}^{\beta_q}([0,1]^{t_q}, K_q(2K_{q-1})^{\beta_q})$ for all j. Having defined these functions, we can now state Lemma 3 of Schmidt-Hieber (2020).

Lemma 3. *Let $(h_{ij})_{i,j}$ be as above with $K_i \geq 1$ for all i. For each $i \in [q]_0$, let*

$$\tilde{h}_i := (\tilde{h}_{ij})_{j=1}^{d_{i+1}} : [0,1]^{d_i} \to \begin{cases} [0,1]^{d_{i+1}} & \textit{if } i < q \\ \mathbb{R} & \textit{if } i = q \end{cases}$$

be a function such that each $\tilde{h}_{ij}$ depends only on t_i variables. Then

$$||h_q \circ .... \circ h_0 - \tilde{h}_q \circ ... \circ \tilde{h}_0||_{L^\infty([0,1]^{d_0})} \leq K_q \left(\prod_{\ell=0}^{q-1} (2K_\ell)^{\beta_{\ell+1}} \right) \sum_{i=0}^{q} |||h_i - \tilde{h}_i|_\infty||_{L^\infty([0,1]^{d_i})}^{\prod_{\ell=i+1}^{q} \beta_\ell \wedge 1}.$$

This lemma is not true. As a counterexample, consider the case where q, d_0, d_1, t_0, t_1, K_0, K_1, β_0 and β_1 are all equal to 1. Let $g_0 = g_1 \equiv 0$. Then $h_0 \equiv 1/2$ and $h_1 \equiv 0$. Define $\tilde{h}_0 \equiv 3/5$ and for $x \in [0,1]$, define

$$\tilde{h}_1(x) := \begin{cases} 1 & \text{if } x = 3/5 \\ 0 & \text{else.} \end{cases}$$

Then $||h_1 \circ h_0 - \tilde{h}_1 \circ \tilde{h}_0||_{L^\infty([0,1])} = 1$, whereas

$$K_1(2K_0)^{\beta_1} \left(||h_0 - \tilde{h}_0||_{L^\infty([0,1])} + ||h_1 - \tilde{h}_1||_{L^\infty([0,1])} \right) = 1/5 < 1.$$

Hence, Lemma 3 does not hold. We assumed here that Schmidt-Hieber (2020) means the

L^∞-norm with respect to Lebesgue measure, but he does not explicitly say this. However, our counterexample easily generalizes to any non-zero measure with a non-empty null set.

The proof of Schmidt-Hieber (2020) seems to assume equality of $||\cdot||_{L^\infty([0,1]^{d_i})}$ and $||\cdot||_\infty$, but this is not true in general. We therefore prove a slightly different version of this lemma, with $||\cdot||_{L^\infty([0,1]^{d_i})}$ replaced by $||\cdot||_\infty$. The proof strategy is the same as that of Schmidt-Hieber (2020), except for replacing the considered norm. This alternative result is sufficient for proving Theorem 1′, since we will only be applying it to continuous functions on domains of the form $[a, b]^t$.

Lemma 3′. *Lemma 3 holds if we replace the norm* $||\cdot||_{L^\infty([0,1]^{d_i})}$ *by the norm* $||\cdot||_\infty$ *on* $[0,1]^{d_i}$ *for all* $i \in [q]_0$.

Proof. Denote $H_i := h_i \circ ... \circ h_0$ and $\tilde{H}_i := \tilde{h}_i \circ ... \circ \tilde{h}_0$, for each i. Further define

$$Q_i := \begin{cases} 1 & \text{if } i = 0 \\ (2K_{i-1})^{\beta_i} & \text{if } i \in [q-1] \\ K_q(2K_{q-1})^{\beta_q} & \text{if } i = q. \end{cases}$$

That is, for each i, Q_i is an upper bound on the Hölder norms of $(h_{ij})_j$. Then for any i and any $\boldsymbol{x} \in \mathbb{R}^{d_0}$, we can apply the triangle inequality as follows:

$$\begin{aligned} |H_i(\boldsymbol{x}) - \tilde{H}_i(\boldsymbol{x})|_\infty &= |h_i \circ H_{i-1}(\boldsymbol{x}) - \tilde{h}_i \circ \tilde{H}_{i-1}(\boldsymbol{x})|_\infty \\ &\le |h_i \circ H_{i-1}(\boldsymbol{x}) - h_i \circ \tilde{H}_{i-1}(\boldsymbol{x})|_\infty + |h_i \circ \tilde{H}_{i-1}(\boldsymbol{x}) - \tilde{h}_i \circ \tilde{H}_{i-1}(\boldsymbol{x})|_\infty \\ &\le Q_i |H_{i-1}(\boldsymbol{x}) - \tilde{H}_{i-1}(\boldsymbol{x})|_\infty^{\beta_i \wedge 1} + |||h_i - \tilde{h}_i|_\infty||_\infty. \end{aligned}$$

Note that we can apply this same bound to the first term of the final line. We then obtain

$$\begin{aligned} &|H_i(\boldsymbol{x}) - \tilde{H}_i(\boldsymbol{x})|_\infty \\ &\quad \le Q_i \Big(Q_{i-1}|H_{i-2}(\boldsymbol{x}) - \tilde{H}_{i-2}(\boldsymbol{x})|_\infty^{\beta_{i-1}\wedge 1} + |||h_{i-1} - \tilde{h}_{i-1}|_\infty||_\infty\Big)^{\beta_i \wedge 1} + |||h_i - \tilde{h}_i|_\infty||_\infty. \end{aligned}$$

We combine this with the following inequality:

$$(y + z)^\alpha \le y^\alpha + z^\alpha, \quad \forall y, z \ge 0,\ \alpha \in [0, 1].$$

This gives

$$\begin{aligned}
&|H_i(\boldsymbol{x}) - \tilde{H}_i(\boldsymbol{x})|_\infty \\
&\quad \leq Q_i \left(Q_{i-1} |H_{i-2}(\boldsymbol{x}) - \tilde{H}_{i-2}(\boldsymbol{x})|_\infty^{(\beta_i \wedge 1)(\beta_{i-1} \wedge 1)} + |||h_{i-1} - \tilde{h}_{i-1}|_\infty||_\infty^{\beta_i \wedge 1} \right) + |||h_i - \tilde{h}_i|_\infty||_\infty \\
&\quad = Q_i Q_{i-1} |H_{i-2}(\boldsymbol{x}) - \tilde{H}_{i-2}(\boldsymbol{x})|_\infty^{\beta_i \beta_{i-1} \wedge 1} + Q_i |||h_{i-1} - \tilde{h}_{i-1}|_\infty||_\infty^{\beta_i \wedge 1} + |||h_i - \tilde{h}_i|_\infty||_\infty,
\end{aligned}$$

where we also used that $Q_i \geq 1$. Starting from $i = q$ and applying this argument q times, we obtain

$$\begin{aligned}
|H_q(\boldsymbol{x}) - \tilde{H}_q(\boldsymbol{x})|_\infty &\leq \sum_{i=0}^{q} \left(\prod_{\ell=i+1}^{q} Q_\ell \right) |||h_i - \tilde{h}_i|_\infty||_\infty^{\prod_{\ell=i+1}^{q} \beta_\ell \wedge 1} \\
&\leq \left(\prod_{i=1}^{q} Q_i \right) \sum_{i=0}^{q} |||h_i - \tilde{h}_i|_\infty||_\infty^{\prod_{\ell=i+1}^{q} \beta_\ell \wedge 1} \\
&= K_q \left(\prod_{i=0}^{q-1} (2K_i)^{\beta_{i+1}} \right) \sum_{i=0}^{q} |||h_i - \tilde{h}_i|_\infty||_\infty^{\prod_{\ell=i+1}^{q} \beta_\ell \wedge 1},
\end{aligned}$$

where the second inequality follows from the fact that $Q_i \geq 1$ for all i. This completes the proof. □

We will combine Theorem 2′, Theorem 5 and Lemma 3′ to prove Theorem 1′. We again follow the proof strategy of Schmidt-Hieber (2020), with some minor modifications to solve the problems described in Section 2. We note that replacing Theorem 2 and Lemma 3 by Theorem 2′ and Lemma 3′ does not at all affect the proof.

Proof of Theorem 1′. We will make use of the following properties of NNs. They all follow straightforwardly from (3) and thinking of NNs as layered and directed acyclic graphs.

- If $f_1 \in \mathcal{F}(L_1, \boldsymbol{p}^{(1)}, s_1, 1)$ and $f_2 \in \mathcal{F}_0(L_2, \boldsymbol{p}^{(2)}, s_2, F_2)$ are NNs with $p_{L_1+1}^{(1)} = p_0^{(2)}$, then they can be composed as NNs to obtain an NN

  $$f_2 \circ \sigma_0(f_1) \in \mathcal{F}\left(L_1 + L_2 + 1, \left(p_0^{(1)}, ..., p_{L_1+1}^{(1)}, p_1^{(2)}, ..., p_{L_2+1}^{(2)}\right), s_1 + s_2, F_2\right).$$

- If $f_1 \in \mathcal{F}(L_0, \boldsymbol{p}^{(1)}, s_1, F_0)$ and $f_2 \in \mathcal{F}(L_0, \boldsymbol{p}^{(2)}, s_2, F_0)$ are NNs with $p_0^{(1)} = p_0^{(2)}$, then they can be parallelized to obtain an NN

$$\left[\boldsymbol{x} \mapsto (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}))\right] \in \mathcal{F}\left((L_0, \left(p_0^{(1)}, p_1^{(1)} + p_1^{(2)}, ..., p_{L_0+1}^{(1)} + p_{L_0+1}^{(2)}\right), s_1 + s_2, F_0\right).$$

- Suppose $(L_1, \boldsymbol{p}^{(1)}), (L_2, \boldsymbol{p}^{(2)})$ are network architectures with $L_1 \leq L_2$, $p_0^{(1)} = p_0^{(2)}$, $p_{L_1+1}^{(1)} = p_{L_2+1}^{(2)}$ and $\max_{i\in[L_1]_0} p_i^{(1)} \leq \min_{i\in[L_2]} p_i^{(2)}$. Suppose further that $s_1 + (L_2 - L_1)p_0^{(1)} \leq s_2$. Then $\mathcal{F}(L_1, \boldsymbol{p}^{(1)}, s_1, \infty) \subset \mathcal{F}(L_2, \boldsymbol{p}^{(2)}, s_2, \infty)$. To see this, suppose $f \in \mathcal{F}(L_1, \boldsymbol{p}^{(1)}, s_1, \infty)$ is an NN with a representation as in (3). Then also $f \in \mathcal{F}(L_2, \boldsymbol{p}^{(2)}, s_2, \infty)$, since f can be written as

$$f = W_L \sigma_{\boldsymbol{v}_{L_1}} W_{L_1-1} \sigma_{\boldsymbol{v}_{L-1}} ... W_1 \sigma_{\boldsymbol{v}_1} W_0 \underbrace{\sigma_0 I ... \sigma_0 I}_{L_2 - L_1 \text{ times}},$$

where I is the $(p_0^{(1)} \times p_0^{(1)})$ identity matrix, which has $p_0^{(1)}$ non-zero entries.

We now prove Theorem 1′. It suffices to prove the result for n sufficiently large. That is, for all $n \geq n_0$, where $n_0 \in \mathbb{N}$ is a constant depending only on

$$\Xi := \left(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F, C_{(ii)}, C_{(iii)}, C_{(iv)}^{-}, C_{(iv)}^{+}\right).$$

Throughout the proof, C' will denote a generic positive constant that may change value from line to line and that depends only on Ξ. We first prove the lower bound in (6). Applying Theorem 2′ with $\epsilon = 1/2$ gives

$$R(\hat{f}_n, f_0) \geq \frac{1}{4}\Delta_n(\hat{f}_n, f_0) - C' \frac{(s+1)\log\left(n(s+1)^L d\right)}{n}. \tag{19}$$

Now note that by Assumptions (iv) and (ii), and by the fact that $n\phi_n \log n \leq C'n$, we have for $n \geq d$ that

$$\begin{aligned}(s+1)\log\left(n(s+1)^L d\right)/n &\leq C' n\phi_n \log(n) \log\left(n\left(n\phi_n \log(n)\right)^L d\right)/n \\ &= C'\phi_n \log(n)\left(\log(n) + L\log\left(n\phi_n \log(n)\right) + \log(d)\right) \\ &\leq C'\phi_n \log(n) L \log(n).\end{aligned}$$

Combining this with (19) then gives

$$R(\hat{f}_n, f_0) \geq \frac{1}{4}\Delta_n(\hat{f}_n, f_0) - C_1\phi_n L \log^2(n),$$

for some constant $C_1 > 0$ that depends only on Ξ. Then for C large enough, for example $C > 8C_1 \vee 8$, we have that if $\Delta_n(\hat{f}_n, f_0) \geq C\phi_n L \log^2(n)$, then

$$R(\hat{f}_n, f_0) \geq \frac{1}{4}\Delta_n(\hat{f}_n, f_0) - C_1\phi_n L \log^2(n) \geq \frac{1}{8}\Delta_n(\hat{f}_n, f_0) \geq \frac{1}{C}\Delta_n(\hat{f}_n, f_0),$$

as desired.

To prove the upper bounds in (6) and (5), we write $f = h_q \circ ... \circ h_0$ with the h_i constructed as described above the statement of Lemma 3. Recall that $h_{0j} \in \mathcal{C}_{t_0}^{\beta_0}([0,1]^{t_0}, 1)$ for all j, $h_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([0,1]^{t_i}, (2K)^{\beta_i})$ for all $i \in [q-1]$ and j, and $h_{qj} \in \mathcal{C}_{t_q}^{\beta_q}([0,1]^{t_q}, K(2K)^{\beta_q})$ for all j.

Next, we apply Theorem 5 to each function h_{ij} individually. As noted in Section 2, we were unable to follow how Schmidt-Hieber (2020) derives his (26). To avoid this problem, we apply Theorem 5 in a slightly different way than he does. In particular, unlike Schmidt-Hieber (2020), we let the choice of m and N depend on i. For each $i \in [q]_0$, we define

$$N_i := \lceil cn^{t_i/(2\beta_i^* + t_i)} \rceil, \quad \kappa_i := \frac{\beta_i + t_i}{2\beta_i^* + t_i} \quad \text{and} \quad m_i := \lceil \kappa_i \log_2 n \rceil,$$

where $c > 0$ is a sufficiently small constant that is to be determined. In particular, c will be chosen depending only on Ξ. Let Q_i be defined as in the proof of Lemma 3′, but with $K_\ell = K$ for all ℓ. Note that this choice of N_i satisfies the requirement $N_i \geq (\beta_i + 1)^{t_i} \vee (Q_i + 1)e^{t_i}$ in Theorem 5 for large enough n, since $N_i \to \infty$ as $n \to \infty$, whereas the RHS remains constant. We further define

$$L_i' := 8 + (m_i + 5)(1 + \lceil \log_2(t_i \vee \beta_i) \rceil) \quad \text{and} \quad s_i := \lfloor 141(t_i + \beta_i + 1)^{3+t_i} N_i(m_i + 6) \rfloor.$$

Then for each h_{ij}, applying Theorem 5 yields a function

$$\tilde{h}_{ij} \in \mathcal{F}_0\left(L_i', (t_i, 6(t_i + \lceil \beta_i \rceil)N_i, ..., 6(t_i + \lceil \beta_i \rceil)N_i, 1), s_i, \infty\right),$$

which satisfies

$$\begin{aligned} ||\tilde{h}_{ij} - h_{ij}||_{L^\infty([0,1]^{t_i})} &\leq (2Q_i + 1)(1 + t_i^2 + \beta_i^2)6^{t_i} N_i n^{-\kappa_i} + Q_i 3^{\beta_i} N_i^{-\beta_i/t_i} \\ &\leq C' n^{\frac{t_i}{2\beta_i^* + t_i}} \cdot n^{-\frac{\beta_i + t_i}{2\beta_i^* + t_i}} + C' \left(n^{\frac{t_i}{2\beta_i^* + t_i}}\right)^{-\beta_i/t_i} \\ &\leq C' n^{-\frac{\beta_i}{2\beta_i^* + t_i}}. \end{aligned} \tag{20}$$

Note that $|| \cdot ||_{L^\infty([0,1]^{t_i})}$ on the LHS of the first line can be replaced by $|| \cdot ||_\infty$, since these norms are equal for continuous functions on $[0,1]^{t_i}$. For all $i < q$ and all j, apply to the output of $\tilde{h}_{ij}$ the two additional layers $x \mapsto 1 - \sigma(1 - x)$ to obtain a network h_{ij}^*. That is, $h_{ij}^* = \tilde{h}_{ij} \wedge 1$. Note that these two additional layers each have width 1 and together require 4 additional non-zero parameters. Then we obtain that

$$h_{ij}^* \in \mathcal{F}_0\left(L_i' + 2, (t_i, 6(t_i + \lceil \beta_i \rceil)N_i, ..., 6(t_i + \lceil \beta_i \rceil)N_i, 1), s_i + 4, \infty\right).$$

Recall that h_{ij} has codomain $[0,1]$ and note that $\sigma(h_{ij}^*) = \left(\tilde{h}_{ij} \wedge 1\right) \vee 0$, which means that $\sigma(h_{ij}^*)$ is the projection of $\tilde{h}_{ij}$ onto $[0,1]$. Hence,

$$||\sigma(h_{ij}^*) - h_{ij}||_\infty \leq ||\tilde{h}_{ij} - h_{ij}||_\infty. \tag{21}$$

Now compute the networks $(h_{ij}^*)_{j=1}^{d_{i+1}}$ in parallel and denote the resulting network by h_i^*. Define $r_i := d_{i+1}(t_i + \lceil \beta_i \rceil)$. Then we have that

$$h_i^* \in \mathcal{F}_0\left(L_i' + 2, (d_i, 6r_iN_i, ..., 6r_iN_i, 1), d_{i+1}(s_i + 4), \infty\right).$$

Remark A.5. *Schmidt-Hieber (2020) defines* $r_i := \max_i d_{i+1}(t_i + \lceil \beta_i \rceil)$*, but we assume that* $\max_i$ *should have been omitted. Alternatively, he might have meant to omit the subscript* i *on the LHS.*

Next, define $r := \max_i r_i, N := \max_i N_i$ and consider the network

$$f^* := \tilde{h}_{q1} \circ \sigma_0(h^*_{q-1}) \circ ... \circ \sigma_0(h^*_0) \in \mathcal{F}_0 \left(E, (d, 6rN, ..., 6rN, 1), \sum_{i=0}^{q} d_{i+1}(s_i + 4), \infty \right),$$

where $E := 3q + \sum_{i=0}^{q} L'_i$.

Remark A.6. *Schmidt-Hieber (2020) instead defines* $E := 3(q-1) + \sum_{i=0}^{q} L'_i$*, but we suspect this is an error.*

Now define $A_n := E - \log_2(n) \sum_{i=0}^{q} \kappa_i \left(\lceil \log_2(t_i \vee \beta_i) \rceil + 1 \right)$. Note that by the definition of L'_i, we have that

$$\begin{aligned}
\sum_{i=0}^{q} L'_i - \log_2(n) \sum_{i=0}^{q} \kappa_i \left(\lceil \log_2(t_i \vee \beta_i) \rceil + 1 \right) \\
&= 8(q+1) + \sum_{i=0}^{q} \left(5 + \lceil \kappa_i \log_2 n \rceil - \kappa_i \log_2 n \right) \left(\lceil \log_2(t_i \vee \beta_i) \rceil + 1 \right) \\
&\leq 8(q+1) + 6 \sum_{i=0}^{q} \left(\lceil \log_2(t_i \vee \beta_i) \rceil + 1 \right),
\end{aligned}$$

since $\lceil \kappa_i \log_2 n \rceil - \kappa_i \log_2 n$ is bounded from above by 1. Hence, A_n is bounded in n. Then since $\lceil \log_2(t_i \vee \beta_i) \rceil < \log_2(t_i \vee \beta_i) + 1$, we obtain for n large enough that

$$\begin{aligned}
E &= A_n + \log_2(n) \sum_{i=0}^{q} \kappa_i \left(\lceil \log_2(t_i \vee \beta_i) \rceil + 1 \right) \\
&< \log_2(n) \sum_{i=0}^{q} \kappa_i \left(\log_2(t_i \vee \beta_i) + 2 \right) \\
&= \log_2(n) \sum_{i=0}^{q} \kappa_i \log_2(4t_i \vee 4\beta_i) \leq L,
\end{aligned}$$

where the final inequality follows from (ii') in Theorem 1'.

We now show that if $c > 0$ is chosen appropriately, we have the following embedding for n large enough:

$$\mathcal{F}_0 \left(E, (d, 6rN, ..., 6rN, 1), \sum_{i=0}^{q} d_{i+1}(s_i + 4), \infty \right) \subset \mathcal{F}_0 \left(L, \boldsymbol{p}, s, \infty \right).$$

We already showed that $E \leq L$. Recall that r is a constant depending only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}$.

Then, by (iii) of Theorem 1, we have that

$$\begin{aligned}6r \max_{i\in[q]_0} n^{t_i/(2\beta_i^*+t_i)} &\leq C' \max_{i\in[q]_0} n^{t_i/(2\beta_i^*+t_i)} \\ &= C'n \max_{i\in[q]_0} n^{-2\beta_i^*/(2\beta_i^*+t_i)} \\ &= C'n\phi_n \\ &\leq C' \min_j p_j.\end{aligned}$$

Since C' denotes a constant that only depends on Ξ, we can also choose a sufficiently small constant c that depends only on Ξ and such that $6rN \leq \min_j p_j$ as desired. Note that since $n\phi_n \to \infty$, (iii) of Theorem 1 also guarantees that $\min_j p_j \geq d$ for n large enough. For the sparsity, we must show that

$$\sum_{i=0}^{q} d_{i+1}(s_i+4) + (L-E)d \leq s. \tag{22}$$

Note that by (iii) and (iv) of Theorem 1, we have for n large enough that

$$d(L-E) \quad \leq \quad dL \quad \leq \quad C'n\phi_n \quad \leq \quad C'\frac{s}{\log n} \quad \leq \quad s/2.$$

By (iv) of Theorem 1, we further have that

$$\begin{aligned}\sum_{i=0}^{q} d_{i+1}(s_i+4) &\leq \sum_{i=0}^{q} d_{i+1}\left(C'N(\lceil \kappa_i \log_2 n\rceil + 6) + 4\right) \\ &\leq C'N\log_2(n) \\ &\leq C'c\log_2(n)\max_i n^{t_i/(2\beta_i^*+t_i)} \\ &= C'c\log_2(n)n\phi_n \\ &\leq C'cs.\end{aligned}$$

Hence, we can choose a sufficiently small c that depends only on Ξ such that $\sum_{i=0}^{q} d_{i+1}(s_i+4) \leq s/2$. Together with $d(L-E) \leq s/2$ for n large enough, this then implies that (22) holds for n large enough. This completes the proof of the embedding. This embedding then implies that $f^* \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$.

We now apply Lemma 3′ and combine this with (20), with sup-norm instead of L^∞-norm, and (21). Recall that c is a constant chosen depending only on Ξ, and hence we now allow the generic constants C' to depend on c. We then obtain for n large enough that

$$\begin{aligned}
||f^* - f_0||_\infty &\leq K \left(\prod_{\ell=0}^{q-1} (2K)^{\beta_{\ell+1}} \right) \left(\sum_{i=0}^{q} |||h_i - \sigma(h_i^*)|_\infty||_\infty^{\prod_{\ell=i+1}^{q} \beta_\ell \wedge 1} \right) \\
&\leq C' \sum_{i=0}^{q} |||h_i - \tilde{h}_i|_\infty||_\infty^{\beta_i^*/\beta_i} \\
&\leq C' \sum_{i=0}^{q} \left(n^{-\frac{\beta_i}{2\beta_i^* + t_i}} \right)^{\beta_i^*/\beta_i} \\
&= C' \sum_{i=0}^{q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \\
&\leq C' \max_{i \in [q]_0} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \\
&= C' \sqrt{\phi_n}.
\end{aligned}$$

Since $f^* \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$, this implies that

$$\inf_{f \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty)} ||f - f_0||_\infty^2 \leq C' \phi_n, \tag{23}$$

for n large enough. To be able to combine this with Theorem 2′, we must replace $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ in the infimum by $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$. Note that (23) implies that there exists a sequence $(\tilde{f}_n)_n$ of functions with $\tilde{f}_n \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ such that for sufficiently large n we have that $||\tilde{f}_n - f_0||_\infty^2 \leq C' \phi_n$, where this constant C' is twice the constant C' in (23). Now define

$$f_n^* := \tilde{f}_n \cdot \left(\frac{||f_0||_\infty}{||\tilde{f}_n||_\infty} \wedge 1 \right).$$

Then by (i) of Theorem 1,

$$||f_n^*||_\infty \quad \leq \quad ||f_0||_\infty \quad \leq \quad ||g_q||_\infty \quad \leq \quad K \quad \leq \quad F.$$

We then have $f_n^* \in \mathcal{F}_0(L, \boldsymbol{p}, s, F)$, since f_n^* can be realized as a network by taking the network $\tilde{f}_n$ and multiplying all weights in the final layer by $\frac{||f_0||_\infty}{||\tilde{f}_n||_\infty} \wedge 1$, which requires neither additional depth nor additional non-zero parameters. Next, note that if $||\tilde{f}_n||_\infty \leq ||f_0||_\infty$, then $f_n^* = \tilde{f}_n$ and $||f_n^* - \tilde{f}_n||_\infty = 0$. If $||f_0||_\infty \leq ||\tilde{f}_n||_\infty$, we have by the reverse triangle inequality

$$\begin{aligned}||f_n^* - \tilde{f}_n||_\infty &= ||\tilde{f}_n||_\infty \left(1 - \frac{||f_0||_\infty}{||\tilde{f}_n||_\infty}\right) \\ &= ||\tilde{f}_n||_\infty - ||f_0||_\infty \\ &\leq ||\tilde{f}_n - f_0||_\infty.\end{aligned}$$

Hence, in either case, we have that $||f_n^* - \tilde{f}_n||_\infty \leq ||\tilde{f}_n - f_0||$, which by the triangle inequality then implies for n sufficiently large that

$$\begin{aligned}||f_n^* - f_0||_\infty &= ||f_n^* - \tilde{f}_n + \tilde{f}_n - f_0||_\infty \\ &\leq ||f_n^* - \tilde{f}_n||_\infty + ||\tilde{f}_n - f_0||_\infty \\ &\leq 2||\tilde{f}_n - f_0||_\infty \\ &\leq C'\sqrt{\phi_n}.\end{aligned}$$

Hence, we indeed have for n sufficiently large that

$$\inf_{f \in \mathcal{F}_0(L, \boldsymbol{p}, s, F)} ||f - f_0||_\infty^2 \leq C'\phi_n. \tag{24}$$

Note that

$$n\phi_n \log n \quad \leq \quad C' \max_{i \in [q]_0} n^{\frac{t_i}{2\beta_i^* + t_i}} \log n \quad \leq \quad C'n,$$

since $\beta_i^* > 0$ for all i. Now apply Theorem $2'$ with $\epsilon = 1$. Then by (iv) of Theorem 1 and the fact that $n\phi_n \log n \leq C'n$, we have that

$$\begin{aligned}\tau_{1,n} &\leq C' \frac{(s+1)\log(n(s+1)^L)}{n} \\ &\leq C' \frac{s}{n} \left(L \log(s+1) + \log(n)\right) \\ &\leq C'\phi_n \log(n) \left(L \log(n\phi_n \log n) + \log(n)\right) \\ &\leq C'\phi_n L \log^2 n.\end{aligned}$$

Hence, we obtain

$$
\begin{aligned}
R(\hat{f}_n, f_0) &\leq 4 \inf_{f \in \mathcal{F}_0(L, \boldsymbol{p}, s, F)} ||f - f_0||_\infty^2 + 4\Delta_n(\hat{f}_n, f_0) + \tau_{1,n} \\
&\leq C'\phi_n + 4\Delta_n(\hat{f}_n, f_0) + C'\phi_n L \log^2 n \\
&\leq C'\phi_n L \log^2(n) + 4\Delta_n(\hat{f}_n, f_0).
\end{aligned}
$$

Now let $C > 0$ be any constant depending only on Ξ, and larger than the value of C' in the final line of the previous display. If $\Delta_n(\hat{f}_n, f_0) \leq C\phi_n L \log^2 n$, then we obtain the upper bound in (5). If $C\phi_n L \log^2 n \leq \Delta_n(\hat{f}_n, f_0)$, we instead obtain the upper bound in (6). This completes the proof.

Remark A.7. *Note that this proof implies that any choice of $C > 0$ in Theorem 1 is valid for some sufficiently large choice of $C' > 0$.*

□

Next, we give a detailed version of the proof given by Schmidt-Hieber (2020) for Theorem 3. This proof does not involve any of the other theorems or lemmas of Schmidt-Hieber (2020). Instead, the proof is based on applying Theorem 2.7 of Tsybakov (2009) and the Varshamov-Gilbert bound, see for example Lemma 2.9 in Tsybakov (2009).

Proof of Theorem 3. Let $|| \cdot ||_2$ denote the L^2-norm on $[0,1]^d$ with respect to Lebesgue measure. Let χ be a Lebesgue density for the distribution of $\boldsymbol{X}_1$ on $[0,1]^d$ such that χ is lower bounded by a positive constant γ and upper bounded by a positive constant Γ. Such a density exists by assumption.

Let $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ be as in (2), with $d_0 = d$ and $t_i \leq \min_{j \in [i-1]_0} d_j$ for all i. Let $\hat{f}_n$ be an estimator taking value in the space of measurable functions $[0,1]^d \to \mathbb{R}$. For any measurable function $f : [0,1]^d \to \mathbb{R}$, let P_f denote the joint distribution of a single observation $(\boldsymbol{X}_1, Y_1)$ under the model (1) with $f_0 = f$ and let P_f^n denote the joint distribution of the whole sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$. Recall that $R(\hat{f}_n, f_0)$ is defined as

$$
R(\hat{f}_n, f_0) := \mathbb{E}_{f_0}\left[\left(\hat{f}_n(\boldsymbol{X}) - f_0(\boldsymbol{X})\right)^2\right],
$$

where $\boldsymbol{X}$ has the same distribution as $\boldsymbol{X}_1$ and is independent of $(\boldsymbol{X}_i, \epsilon_i)_{i=1}^n$. The latter condition implies that $\boldsymbol{X}$ is also independent of $(\boldsymbol{X}_i, Y_i)_{i=1}^n$. To emphasize that $\hat{f}_n$ depends on $(\boldsymbol{X}_i, Y_i)_{i=1}^n$, we write $\hat{f}_n^{(\boldsymbol{X}_i,Y_i)_{i=1}^n} := \hat{f}_n$. Then we have by Fubini's theorem that

$$\begin{aligned} R(\hat{f}_n, f_0) &= \int_{\left([0,1]^d \times \mathbb{R}\right)^n} \int_{[0,1]^d} \left(\hat{f}_n^{\boldsymbol{I}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\right)^2 \chi(\boldsymbol{x}) d\boldsymbol{x} dP_{f_0}^n(\boldsymbol{I}) \\ &\geq \gamma \int_{\left([0,1]^d \times \mathbb{R}\right)^n} \int_{[0,1]^d} \left(\hat{f}_n^{\boldsymbol{I}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\right)^2 d\boldsymbol{x} dP_{f_0}^n(\boldsymbol{I}) \\ &= \gamma \int_{\left([0,1]^d \times \mathbb{R}\right)^n} ||\hat{f}_n^{\boldsymbol{I}} - f_0||_2^2 d\boldsymbol{x} dP_{f_0}^n(\boldsymbol{I}) \\ &= \gamma \mathbb{E}_{f_0} \left[||\hat{f}_n^{(\boldsymbol{X}_i,Y_i)_{i=1}^n} - f_0||_2^2 \right]. \end{aligned}$$

The same approach can be used to show that $R(\hat{f}_n, f_0) \leq \Gamma \mathbb{E}_{f_0} \left[||\hat{f}_n^{(\boldsymbol{X}_i,Y_i)_{i=1}^n} - f_0||_2^2 \right]$.

Let D_1, D_2 be two probability measures on the same measurable space and denote the joint distribution of an iid sample of size n from the distribution D_i by D_i^n. Then it is well-known that the Kullback-Leibler divergence satisfies $\mathrm{KL}(D_1^n, D_2^n) = n\,\mathrm{KL}(D_1, D_2)$. Write N_{μ,σ^2} for the law of a normal random variable with mean μ and variance σ^2 and $p_{N(\mu,\sigma^2)}$ for its Lebesgue density. Then it is also well-known that $\mathrm{KL}(N_{\mu_1,1}, N_{\mu_2,1}) = (\mu_1 - \mu_2)^2/2$. Now let $f : [0,1]^d \to \mathbb{R}$ be a measurable function. Then the probability distribution P_f has Lebesgue density $p_{f;(\boldsymbol{X},Y)}$ given by

$$\begin{aligned} p_{f;(\boldsymbol{X},Y)}(\boldsymbol{x}, y) &= p_{f;(Y|\boldsymbol{X})}(y|\boldsymbol{x}) \cdot p_{f;\boldsymbol{X}}(\boldsymbol{x}) \\ &= p_{N(f(\boldsymbol{x}),1)}(y) \cdot \chi(\boldsymbol{x}), \end{aligned}$$

where we use that $Y|\boldsymbol{X} \sim N(f(\boldsymbol{X}), 1)$. Now note that if $g : [0,1]^d \to \mathbb{R}$ is also a measurable function, then those properties imply that the Kullback-Leibler divergence between P_f^n and

P_g^n satisfies

$$\begin{aligned}
\mathrm{KL}(P_f^n, P_g^n) &= n\,\mathrm{KL}(P_f, P_g) \\
&= n \int_{[0,1]^d} \int_{\mathbb{R}} \log\left(\frac{p_{f;(\boldsymbol{X},Y)}(\boldsymbol{x},y)}{p_{g;(\boldsymbol{X},Y)}(\boldsymbol{x},y)}\right) p_{f;(\boldsymbol{X},Y)}(\boldsymbol{x},y) dy d\boldsymbol{x} \\
&= n \int_{[0,1]^d} \int_{\mathbb{R}} \log\left(\frac{p_{N(f(\boldsymbol{x}),1)}(y)}{p_{N(g(\boldsymbol{x}),1)}(y)}\right) p_{N(f(\boldsymbol{x}),1)}(y) dy \chi(\boldsymbol{x}) d\boldsymbol{x} \\
&= n \int_{[0,1]^d} \mathrm{KL}(N_{f(\boldsymbol{x},1}, N_{g(\boldsymbol{x}),1}) \chi(\boldsymbol{x}) d\boldsymbol{x} \qquad (25) \\
&= \frac{n}{2} \int_{[0,1]^d} (f(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \chi(\boldsymbol{x}) d\boldsymbol{x} \\
&\leq \frac{n\Gamma}{2} \int_{[0,1]^d} (f(\boldsymbol{x}) - g(\boldsymbol{x}))^2 d\boldsymbol{x} \\
&= \frac{n\Gamma}{2} ||f-g||_2^2.
\end{aligned}$$

We now wish to apply Theorem 2.7 of Tsybakov (2009) with $w(x) = x^2$ and $\Theta = \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$, distance $d(f,g) = ||f-g||_2$, $\alpha = 1/18$, $\phi = 1$ and $s = A = \kappa\sqrt{\phi_n}/2$ for some $\kappa > 0$. To do so, we find some $M \in \mathbb{N}$, $\kappa > 0$ and $(\theta_0, ..., \theta_M) = (f_{(0)}, ..., f_{(M)}) \in \Theta^{M+1}$ such that

$$||f_{(i)} - f_{(j)}||_2 \geq \kappa\sqrt{\phi_n} \text{ for all } i, j \in [M]_0 \text{ with } i \neq j, \qquad (26)$$

and

$$\frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(P^n_{f_{(j)}}, P^n_{f_{(0)}}) \leq \frac{1}{18} \log M. \qquad (27)$$

For (27), it is sufficient that the following condition holds:

$$n \sum_{i=1}^{M} ||f_{(i)} - f_{(0)}||_2^2 \leq \frac{M}{9\Gamma} \log(M). \qquad (28)$$

This follows immediately from applying (25) to the LHS of (27).

Note that we omitted the absolute continuity condition in (ii) of Theorem 2.7 of Tsybakov (2009), since the Gaussian noise in our model immediately implies that P_f and P_g are mutually absolutely continuous for any measurable functions $f, g : [0,1]^d \to \mathbb{R}$.

We now choose κ, M and $(f_{(0)}, ..., f_{(M)})$ such that (26) and (28) hold. Fix some $i^* \in$

$\operatorname{argmin}_i \frac{\beta_i^*}{2\beta_i^*+t_i}$ and define $\beta^* := \beta_{i^*}$, $\beta^{**} := \beta_{i^*}^*$ and $t^* := t_{i^*}$. Then $\phi_n = n^{\frac{-2\beta^{**}}{2\beta^{**}+t^*}}$.

Further fix some $H \in \mathcal{C}_1^{\beta^*}(\mathbb{R}, 1)$ with support $[0,1]$. Note that such a function always exists and that $\int_0^1 H(x)^2 dx < \infty$ due to the continuity of H on the compact domain $[0,1]$. Define

$$m_n := \lfloor \rho n^{\frac{1}{2\beta^{**}+t^*}} \rfloor \quad \text{and} \quad h_n := 1/m_n,$$

where $\rho > 0$ is some large enough constant such that $m_n \geq 8$ and

$$nh_n^{2\beta^{**}+t^*} \leq \frac{1}{72\Gamma ||H^B||_2^{2t^*}}, \quad \text{where} \quad B := \prod_{\ell=i^*+1}^{q} (\beta_\ell \wedge 1).$$

Remark A.8. *Schmidt-Hieber (2020) requires that* $nh_n^{2\beta^*+t^*} \leq \frac{1}{72\Gamma||H^B||_2^{2t^*}}$*, instead of* $nh_n^{2\beta^{**}+t^*} \leq \frac{1}{72\Gamma||H^B||_2^{2t^*}}$*. We suspect this is a mistake. Note that the function that we denote by H, Schmidt-Hieber (2020) denotes by K. We use H instead, because K is already used to denote the bound on the Hölder norm in the statement of Theorem 3.*

H will act as a kernel and h_n will act as a bandwidth. Define $\mathcal{U}_n := \{0, h_n, 2h_n, ..., (m_n - 1)h_n\}^{t^*}$ and for any $\boldsymbol{u} = (u_1, ..., u_{t^*}) \in \mathcal{U}_n$, define $\psi_{\boldsymbol{u}} : [0,1]^{t^*} \to \mathbb{R}$ by

$$\psi_{\boldsymbol{u}}(\boldsymbol{x}) := h_n^{\beta^*} \prod_{j=1}^{t^*} H\left(\frac{x_j - u_j}{h_n}\right), \quad \text{where} \quad \boldsymbol{x} = (x_1, ..., x_{t^*}).$$

Since H has Hölder smoothness β^*, $\psi_{\boldsymbol{u}}$ also has Hölder smoothness β^*. We will now bound its Hölder norm. For any $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ with $|\alpha| < \beta^*$, a basic calculation yields that

$$\partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}(\boldsymbol{x}) = h_n^{\beta^*-|\alpha|} \prod_{j=1}^{t^*} H^{(\alpha_j)}\left(\frac{x_j - u_j}{h_n}\right),$$

where $H^{(\alpha_j)}$ denotes the α_j-th derivative of H. Then for any $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ with $|\boldsymbol{\alpha}| < \beta^*$, we have $||\partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}||_\infty \leq 1$, since $h_n \leq 1$ and H has Hölder norm bounded by 1. Recall that for any $a_1, ..., a_{t^*}, b_1, ..., b_{t^*} \in [-1,1]$, we have the inequality $\left|\prod_{j=1}^{t^*} a_j - \prod_{j=1}^{t^*} b_j\right| \leq \sum_{j=1}^{t^*} |a_j - b_j|$. For any $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ with $|\boldsymbol{\alpha}| = \underline{\beta^*}$, this inequality and the fact that $H \in \mathcal{C}_1^{\beta^*}(\mathbb{R}, 1)$ give for all $\boldsymbol{x} = (x_1, ..., x_{t^*}), \boldsymbol{y} = (y_1, ..., y_{t^*})$ with $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^{t^*}$ that

$$\begin{aligned}
\frac{|\partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\overline{\beta^*}}} &= \frac{h_n^{\overline{\beta^*}}}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\overline{\beta^*}}} \cdot \left| \prod_{j=1}^{t^*} H^{(\alpha_j)} \left(\frac{x_j - u_j}{h_n} \right) - \prod_{j=1}^{t^*} H^{(\alpha_j)} \left(\frac{y_j - u_j}{h_n} \right) \right| \\
&\leq \sum_{j=1}^{t^*} h_n^{\overline{\beta^*}} \frac{\left| H^{(\alpha_j)} \left(\frac{x_j - u_j}{h_n} \right) - H^{(\alpha_j)} \left(\frac{y_j - u_j}{h_n} \right) \right|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\overline{\beta^*}}} \qquad (29) \\
&\leq \sum_{j=1}^{t^*} 1 \cdot 1 = t^*.
\end{aligned}$$

Now note that there are $\left(\underline{\beta^*}\right)^{t^*}$ choices of $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ possible with $|\boldsymbol{\alpha}| = \underline{\beta^*}$. Using this fact and the fact that there is only one choice of $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ with $|\boldsymbol{\alpha}| = 0$, it is also straightforward to upper bound the number of choices of $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ such that $|\boldsymbol{\alpha}| < \beta^*$ by $\underline{\beta^*}(\underline{\beta^*})^{t^*} + 1$. Hence, we can upper bound the Hölder norm of $\psi_{\boldsymbol{u}}$ by

$$\beta^*(\beta^*)^{t^*} + 1 + (\beta^*)^{t^*} t^* \quad \leq \quad (\beta^*)^{t^*}(\beta^* + t^*) + 1 \quad =: \quad U.$$

That is, $\psi_{\boldsymbol{u}} \in \mathcal{C}_{t^*}^{\beta^*}([0,1]^{t^*}, U)$. Next, for any $\boldsymbol{w} = (w_{\boldsymbol{u}})_{\boldsymbol{u} \in \mathcal{U}_n} \in \{0,1\}^{\mathcal{U}_n} =: \mathcal{W}_n$, define

$$\phi_{\boldsymbol{w}} := \sum_{\boldsymbol{u} \in \mathcal{U}_n} w_{\boldsymbol{u}} \psi_{\boldsymbol{u}}.$$

Note that by construction, $\psi_{\boldsymbol{u}}$ and $\psi_{\boldsymbol{v}}$ have disjoint support if $\boldsymbol{u} \neq \boldsymbol{v}$. We will use this to bound the Hölder norm of $\phi_{\boldsymbol{w}}$. Let $\boldsymbol{w} \in \mathcal{W}_n$. Because of the disjoint supports of $(\psi_{\boldsymbol{u}})_{\boldsymbol{u} \in \mathcal{U}_n}$, we then have that their partial derivatives also have disjoint support and we obtain for any $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ with $|\alpha| < \beta^*$ that

$$\begin{aligned}
||\partial^{\boldsymbol{\alpha}} \phi_{\boldsymbol{w}}||_\infty &= \sup_{\boldsymbol{x} \in [0,1]^{t^*}} \left| \sum_{\boldsymbol{u} \in \mathcal{U}_n} \partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}(\boldsymbol{x}) \right| \\
&= \sup_{\boldsymbol{x} \in [0,1]^{t^*}} \max_{\boldsymbol{u} \in \mathcal{U}_n : \boldsymbol{w}_{\boldsymbol{u}} = 1} |\partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}(\boldsymbol{x})| \leq 1,
\end{aligned}$$

with the final bound following from the fact that $||\partial^{\boldsymbol{\alpha}} \psi_{\boldsymbol{u}}||_\infty \leq 1$ for any $\boldsymbol{u} \in \mathcal{U}_n$. Now suppose $\boldsymbol{\alpha} \in \mathbb{N}_0^{t^*}$ with $|\boldsymbol{\alpha}| = \underline{\beta^*}$ and $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^{t^*}$ with $\boldsymbol{x} \neq \boldsymbol{y}$. Denote the support of a function f by $\mathrm{supp}(f)$. Note that the bound (29) also holds for $\phi_{\boldsymbol{w}}$ instead of $\psi_{\boldsymbol{u}}$ in all three of the following cases:

- there exists a $\boldsymbol{u} \in \mathcal{U}_n$ with $\boldsymbol{w}_{\boldsymbol{u}} = 1$ such that $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{supp}(\psi_{\boldsymbol{u}})$,

- $\boldsymbol{x} \notin \operatorname{supp}(\phi_{\boldsymbol{w}})$ and there exists a $\boldsymbol{u} \in \mathcal{U}_n$ with $\boldsymbol{w}_{\boldsymbol{u}} = 1$ such that $\boldsymbol{y} \in \operatorname{supp}(\psi_{\boldsymbol{u}})$,

- $\boldsymbol{x}, \boldsymbol{y} \notin \operatorname{supp}(\phi_{\boldsymbol{w}})$.

The final case to consider is if there exist $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}_n$ with $\boldsymbol{u} \neq \boldsymbol{v}$ and $\boldsymbol{w}_{\boldsymbol{u}} = \boldsymbol{w}_{\boldsymbol{v}} = 1$ such that $\boldsymbol{x} \in \operatorname{supp}(\psi_{\boldsymbol{u}}), \boldsymbol{y} \in \operatorname{supp}(\psi_{\boldsymbol{v}})$. In that case, we have that $\boldsymbol{x} \notin \operatorname{supp}(\psi_{\boldsymbol{v}})$ and $\boldsymbol{y} \notin \operatorname{supp}(\psi_{\boldsymbol{u}})$. We then obtain

$$\begin{aligned}\frac{|\partial^{\boldsymbol{\alpha}}\phi_{\boldsymbol{w}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}}\phi_{\boldsymbol{w}}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\overline{\beta^*}}} &= \frac{|\partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{u}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{v}}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\overline{\beta^*}}} \\ &= \frac{|\partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{u}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{u}}(\boldsymbol{y}) + \partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{v}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{v}}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\overline{\beta^*}}} \\ &\leq \frac{|\partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{u}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{u}}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\overline{\beta^*}}} + \frac{|\partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{v}}(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}}\psi_{\boldsymbol{v}}(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|_\infty^{\overline{\beta^*}}} \leq 2t^*,\end{aligned}$$

where the penultimate inequality follows from the triangle inequality and in the final inequality follows from (29). In particular, this shows that the the Hölder norm of $\phi_{\boldsymbol{w}}$ can be bounded by $2U$. That is, $\phi_{\boldsymbol{w}} \in \mathcal{C}_{t^*}^{\beta^*}([0,1]^{t^*}, 2U)$.

Now for $i \in [t^*]_0$ and $\boldsymbol{x} = (x_1, ..., x_{d_i}) \in [0,1]^{d_i}$, define

$$g_i(\boldsymbol{x}) := (x_1, ..., x_{d_{i+1}}) \quad \text{if } d_{i+1} \leq d_i$$

and

$$g_i(\boldsymbol{x}) := (x_1, ..., x_{d_i}, 0, ..., 0) \in \mathbb{R}^{d_{i+1}} \quad \text{if } d_{i+1} > d_i.$$

For $i = i^*$, $\boldsymbol{w} \in \mathcal{W}_n$ and $\boldsymbol{x} = (x_1, ..., x_{d_{i^*}}) \in [0,1]^{d_{i^*}}$, define

$$g_{i^*,\boldsymbol{w}}(\boldsymbol{x}) := (\phi_{\boldsymbol{w}}(x_1, ..., x_{t^*}), 0, ..., 0) \in \mathbb{R}^{d_{i^*+1}}.$$

For $i \in [q] \setminus [i^*]$ and $\boldsymbol{x} = (x_1, ..., x_{d_i}) \in [0,1]^{d_i}$, define

$$g_i(\boldsymbol{x}) = (x_1^{\beta_i \wedge 1}, 0, ..., 0) \in \mathbb{R}^{d_{i+1}}.$$

For any $\boldsymbol{w} \in \mathcal{W}_n$, this allows us to define

$$f_{\boldsymbol{w}}(\boldsymbol{x}) := g_q \circ g_{q-1} \circ ... \circ g_{i^*+1} \circ g_{i^*,\boldsymbol{w}} \circ g_{i^*-1} \circ ... \circ g_0.$$

Further recall that $B := \prod_{\ell=i^*+1}^{q}(\beta_\ell \wedge 1)$ and note that $B\beta^* = \beta^{**}$. Then since $t_{i^*} \leq \min\{d_0, ..., d_{i^*-1}\}$, we have for any $\boldsymbol{x} \in [0,1]^d$ that the first t^* elements of $g_{i^*-1} \circ ... \circ g_0(\boldsymbol{x})$ are $(x_1, ..., x_{t^*})$. Hence, it becomes straightforward to compute that

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \phi_{\boldsymbol{w}}(\boldsymbol{x})^B = \left(\sum_{\boldsymbol{u} \in \mathcal{U}_n} \boldsymbol{w}_{\boldsymbol{u}} \psi_{\boldsymbol{u}}(\boldsymbol{x})\right)^B.$$

Due to all $\psi_{\boldsymbol{u}}$ having disjoint supports for different $\boldsymbol{u}$, this equals

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{\boldsymbol{u} \in \mathcal{U}_n} \boldsymbol{w}_{\boldsymbol{u}} \psi_{\boldsymbol{u}}(\boldsymbol{x})^B.$$

Furthermore, note that each g_i has arbitrarily high Hölder smoothness and that their Hölder semi-norms for any particular smoothness index only depend on $\boldsymbol{d}$ and $\boldsymbol{\beta}$. Since U depends only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}$, we have that $f_{\boldsymbol{w}} \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$, for some sufficiently large K that can be chosen based on only $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}$.

Note that for any $\boldsymbol{u} \in \mathcal{U}_n$, writing $\boldsymbol{x} = (x_1, ..., x_{t^*})$, we have

$$\begin{aligned} ||\psi_{\boldsymbol{u}}^B||_2^2 &= \int_{[0,1]^{t^*}} \left(h_n^{\beta^*} \prod_{j=1}^{t^*} H\left(\frac{x_j - u_j}{h_n}\right)\right)^{2B} d\boldsymbol{x} \\ &= h_n^{2\beta^* *} \prod_{j=1}^{t^*} \left(\int_{u_j}^{u_j + h_n} H\left(\frac{x_j - u_j}{h_n}\right)^{2B} dx_j \right) \\ &= h_n^{2\beta^{**}} \prod_{j=1}^{t^*} \left(h_n ||H^B||_2^2\right) \\ &= h_n^{2\beta^{**} + t^*} ||H^B||_2^{2t^*}, \end{aligned} \tag{30}$$

where the second equality follows from H having support $[0,1]$.

Remark A.9. *Schmidt-Hieber (2020) claims that* $||\psi_{\boldsymbol{u}}||_2^2 = h_n^{2\beta^{**}+t^*}||H^B||_2^{2t^*}$*, but this seems to be an error.*

For $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}_n$, define their Hamming distance to be

$$\operatorname{Ham}(\boldsymbol{w}, \boldsymbol{w}') := \sum_{\boldsymbol{u} \in \mathcal{U}_n} \mathbf{1}(\boldsymbol{w}_{\boldsymbol{u}} \neq \boldsymbol{w}'_{\boldsymbol{u}}).$$

Using (30), we can compute that

$$\begin{aligned} ||f_{\boldsymbol{w}} - f_{\boldsymbol{w}'}||_2^2 &= \int_{[0,1]^d} \left(\sum_{\boldsymbol{u} \in \mathcal{U}_n} (\boldsymbol{w}_{\boldsymbol{u}} - \boldsymbol{w}'_{\boldsymbol{u}})^2 \psi_{\boldsymbol{u}}(\boldsymbol{x})^B \right)^2 d\boldsymbol{x} \\ &= \sum_{\boldsymbol{u} \in \mathcal{U}_n} \mathbf{1}(\boldsymbol{w}_{\boldsymbol{u}} \neq \boldsymbol{w}'_{\boldsymbol{u}}) \int_{[0,1]^d} \psi_{\boldsymbol{u}}(\boldsymbol{x})^{2B} d\boldsymbol{x} \\ &= \sum_{\boldsymbol{u} \in \mathcal{U}_n} \mathbf{1}(\boldsymbol{w}_{\boldsymbol{u}} \neq \boldsymbol{w}'_{\boldsymbol{u}}) ||\psi_{\boldsymbol{u}}^B||_2^2 \\ &= h_n^{2\beta^{**}+t^*} ||H^B||_2^{2t^*} \operatorname{Ham}(\boldsymbol{w}, \boldsymbol{w}'), \end{aligned} \tag{31}$$

where the second equality follows from $\psi_{\boldsymbol{u}}$ having disjoint support for different values of $\boldsymbol{u}$. We now apply the Varshamov-Gilbert bound, see Lemma 2.7 of Tsybakov (2009). We let $m = |\mathcal{U}_n| = m_n^{t^*}$ so that $\Omega = \mathcal{W}_n$ and note that we chose ρ such that $8 \leq m_n \leq m_n^{t^*}$. Then the Varshamov-Gilbert bound guarantees the existence of an integer $M \geq 2^{m_n^{t^*}/8}$ and elements $\boldsymbol{w}^{(0)}, ..., \boldsymbol{w}^{(M)} \in \mathcal{W}_n$ such that $\operatorname{Ham}(\boldsymbol{w}^{(i)}, \boldsymbol{w}^{(j)}) \geq m_n^{t^*}/8$ for all $i, j \in [M]_0$ with $i \neq j$.

Remark A.10. *The symbol ρ used by Tsybakov (2009) denotes Hamming distance and is unrelated to the constant ρ that we defined.*

We choose this M to be the M for Theorem 2.7 of Tsybakov (2009) and we choose $(f_{(0)}, ..., f_{(M)}) = (f_{\boldsymbol{w}^{(0)}}, ..., f_{\boldsymbol{w}^{(M)}})$. We further choose $\kappa = ||H^B||_2^{t^*}/(\sqrt{8}\rho^{\beta^{**}})$. We now show that these choices satisfy (26) and (28).

For any $i, j \in [M]_0$ with $i \neq j$, (31) now implies that

$$\begin{aligned} ||f_{(i)} - f_{(j)}||_2^2 &= h_n^{2\beta^{**}+t^*} ||H^B||_2^{2t^*} \operatorname{Ham}(\boldsymbol{w}^{(i)}, \boldsymbol{w}^{(j)}) \\ &\geq h_n^{2\beta^{**}+t^*} ||H^B||_2^{2t^*} m_n^{t^*}/8 \\ &= (h_n \rho)^{2\beta^{**}} \kappa^2 \\ &\geq n^{\frac{2\beta^{**}}{2\beta^{**}+t^*}} \kappa^2 = \kappa^2 \phi_n, \end{aligned}$$

where we used the definitions of m_n and h_n. This shows that (26) holds.

Similarly, for any $i \in [M]_0$, we can bound the Hamming distance trivially by $\mathrm{Ham}(\boldsymbol{w}, \boldsymbol{w}') \leq |\mathcal{U}_n| = m_n^{t^*}$, for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}_n$. Recall that $nh_n^{2\beta^{**}+t^*} \leq \frac{1}{72\Gamma ||H^B||_2^{2t^*}}$ and note that $M \geq 2^{m_n^{t^*}/8}$ implies that $m_n^{t^*} \leq 8 \log M$. Combining these three bounds, we obtain

$$\begin{aligned} n||f_{(i)} - f_{(0)}||_2^2 &= nh_n^{2\beta^{**}+t^*} ||H^B||_2^{2t^*} \mathrm{Ham}(\boldsymbol{w}^{(i)}, \boldsymbol{w}^{(0)}) \\ &\leq nh_n^{2\beta^{**}+t^*} ||H^B||_2^{2t^*} m_n^{t^*} \\ &\leq \frac{m_n^{t^*}}{72\Gamma} \\ &\leq \frac{\log M}{9\Gamma}. \end{aligned}$$

Hence, (28) is also satisfied. This means that we may now apply Theorem 2.7 of Tsybakov (2009). This implies that there exists an absolute constant λ such that

$$\inf_{\hat{f}_n} \sup_{f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)} R(\hat{f}_n, f_0) \geq \lambda \kappa^2 \phi_n.$$

Choosing $c = \lambda\kappa^2$ then completes the proof of Theorem 3.

□

We can now combine Theorems $2'$ and 3 to prove Lemma 1.

Proof of Lemma 1. For $s = 0$, the result is trivial since $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ consists only of the zero function. Therefore, suppose $s \geq 1$. Denote $\mathcal{C} := \mathcal{C}_d^\beta([0,1]^d, K)$. We will choose $c_2 < 1$. Consider the d-variate nonparametric regression model (1). Suppose $(\boldsymbol{X}_i)_{i=1}^n$ are iid with uniform distributions on $[0,1]^d$. We require that K is large enough to apply Theorem 3 to $\mathcal{G}(0, d, d, \beta, K) = \mathcal{C}$. Then $\phi_n = n^{-2\beta/(2\beta+d)}$. Denote the constant c in the statement of Theorem 3 by c_3. Note that it is sufficient to consider the infimum over only $\mathcal{F}_0(L, \boldsymbol{p}, s, K+1)$ instead of $\mathcal{F}_0(L, \boldsymbol{p}, s, \infty)$ on the LHS of (7), since $||f_0||_\infty \leq K$ and hence $||f_0 - f||_\infty > 1 \geq \epsilon$ for all $f_0 \in \mathcal{C}$ and $f \in \mathcal{F}_0(L, \boldsymbol{p}, s, \infty) \setminus \mathcal{F}(L, \boldsymbol{p}, s, K+1)$. Next, let $\kappa > 0$ be some positive constant and for each n, let $\hat{f}_n$ be an estimator taking values in

$\mathcal{F}_0(L, \boldsymbol{p}, s, K+1)$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_n(\boldsymbol{X}_i))^2 - \inf_{f \in \mathcal{F}_0(L,\boldsymbol{p},s,K+1)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2 \leq \kappa.$$

Then $\Delta_n(\hat{f}_n, f_0) \leq \kappa$ for all $f_0 \in \mathcal{C}$, where $\Delta_n(\hat{f}_n, f_0)$ is as defined in (4) for $F = K+1$. Then by Theorem 3,

$$\sup_{f_0 \in \mathcal{C}} R\left(\hat{f}_n, f_0\right) \quad \geq \quad c_3\phi_n \quad = \quad c_3 n^{-\frac{2\beta}{2\beta+d}}, \tag{32}$$

for all n. Then by Theorem $2'$ with $F = K+1$ and $\epsilon = 1$, we have

$$R(\hat{f}_n, f_0) \leq 4\left(\inf_{f \in \mathcal{F}_0(L,\boldsymbol{p},s,K+1)} ||f - f_0||_\infty^2 + \kappa\right) + C_1(K+1)^4 n^{-1}(s+1)\log(n(s+1)^L d), \tag{33}$$

for any $f_0 \in \mathcal{C}$. Hence, (33) also holds if we take the supremum over $f_0 \in \mathcal{C}$ on both the LHS and RHS. Now let $c_2 \leq \sqrt{c_3/8}$. For any $\epsilon \in (0, c_2]$, define $n_\epsilon := \lfloor(\sqrt{8}\epsilon/\sqrt{c_3})^{-(2\beta+d)/\beta}\rfloor$. Then straightforward calculations yield $n_\epsilon \geq 1$, and $(2n_\epsilon)^{-1} \leq (\sqrt{8}\epsilon/\sqrt{c_3})^{(2\beta+d)/\beta}$, and $n_\epsilon^{-2\beta/(2\beta+d)} \geq 8\epsilon^2/c_3$. Denote by $C' > 0$ a generic constant depending only on K, β, d whose value might change from line to line. Now let $n = n_\epsilon$ and combine the previous bounds with the lower bound of (32) and the upper bound of (33) to obtain

$$8\epsilon^2 \leq 4\sup_{f_0 \in \mathcal{C}} \inf_{f \in \mathcal{F}_0(L,\boldsymbol{p},s,K+1)} ||f - f_0||_\infty^2 + 4\kappa + C'\epsilon^{(2\beta+d)/\beta} s \log(n_\epsilon(s+1)^L d). \tag{34}$$

Since this bounds holds for all $\kappa > 0$, it also holds for $\kappa = 0$. Note further that

$$\log(n_\epsilon(s+1)^L d) \leq C'\left[\log(\epsilon^{-1}) + L\log(s) + L\right].$$

Applying these two observations to (34), we obtain

$$8\epsilon^2 \leq 4\sup_{f_0 \in \mathcal{C}} \inf_{f \in \mathcal{F}_0(L,\boldsymbol{p},s,K+1)} ||f - f_0||_\infty + C'\epsilon^{(2\beta+d)/\beta} s\left[\log(\epsilon^{-1}) + L\log(s) + L\right]. \tag{35}$$

Now suppose that $s \leq c_1(\epsilon^{d/\beta} L \log(1/\epsilon))^{-1}$ for some constant $c_1 > 0$. We choose $c_2 < 1/e$ and $c_1 < 1$ depending only on K, β, d. Then $\log(\epsilon^{-1}) > 1$ and we can upper bound the second term of the upper bound in (35) as follows:

$$\begin{aligned}
\epsilon^{(2\beta+d)/\beta} s \left[\log(\epsilon^{-1}) + L\log(s) + L\right] &\leq c_1\epsilon^2 \frac{\log(\epsilon^{-1}) + L\log\left(c_1\epsilon^{-d/\beta}/\left(L\log(1/\epsilon)\right)\right) + L}{L\log(\epsilon^{-1})} \\
&\leq c_1\epsilon^2 \frac{\log(\epsilon^{-1}) + L\log\left(\epsilon^{-d/\beta}\right) + L}{L\log(\epsilon^{-1})} \\
&\leq c_1\epsilon^2 \frac{\left((1+\frac{d}{b}L\right)\log(\epsilon^{-1}) + L}{L\log(\epsilon^{-1})} \\
&\leq c_1 C'\epsilon^2.
\end{aligned}$$

We choose $c_1 > 0$ sufficiently small depending only on K, β, d, so that the final RHS of this display is bounded from above by $4\epsilon^2$. Returning to (35), we then obtain

$$8\epsilon^2 \leq 4 \sup_{f_0\in\mathcal{C}} \inf_{f\in\mathcal{F}_0(L,\boldsymbol{p},s,K+1)} ||f - f_0||_\infty^2 + 4\epsilon^2.$$

That is, $\epsilon \leq \sup_{f_0\in\mathcal{C}} \inf_{f\in\mathcal{F}_0(L,\boldsymbol{p},s,K+1)} ||f - f_0||_\infty$, as desired.

□

B Proofs of the propositions

Proof of Proposition 1. We perform induction on q. If $q = 0$, the result is trivial. Now suppose the result holds for some $q \in \mathbb{N}_0$ and all $\boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ as in (2). We show it also holds for $q+1$. Let $\boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K$ be as in (2), but with $q+1$ instead of q. Let $f = (f_j)_{j=1}^{d_{q+2}} \in \mathcal{G}(q+1, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$. Let $f = g_{q+1} \circ ... \circ g_0$ with g_i as in (2), but with $q+1$ instead of q. Then for any $j \in [q+2]$, we have $f_j = g_{(q+1)j} \circ (g_q \circ ... \circ g_0)$. We now distinguish two cases:

1. $\min_{i \in [q]_0} \beta_i \leq 1$,

2. $\min_{i \in [q]_0} \beta_i > 1$.

In case 1, define $\alpha := \prod_{i \in [q]_0 : \beta_i \leq 1} \beta_i$. In case 2, define $\alpha := \min_{i \in [q]_0} \beta_i$. Note that in case 1, we have $\alpha \leq 1$. In case 2, we have $\alpha > 1$. By the induction hypothesis, $h_k := g_{qk} \circ ... \circ g_0$ has Hölder smoothness index α for any $k \in [d_{q+1}]$. Note that each h_k is $\min(1, \alpha)$-Hölder, which implies that $h := (h_k)_{k=1}^{d_{q+1}} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{q+1}}$ is $\min(1, \alpha)$-Hölder. Suppose now that $\beta_{q+1} \leq 1$, then $\min_{i \in [q+1]_0} \beta_i \leq 1$ and for any $j \in [q+2]$, $f_j = g_{(q+1)j} \circ h$ is $(\beta_{q+1} \min(1, \alpha))$-Hölder. Next, note that in case 1, we have that $\alpha \leq 1$ so that

$$\begin{aligned} \beta_{q+1} \min(1, \alpha) &= \beta_{q+1} \alpha \\ &= \beta_{q+1} \prod_{i \in [q]_0 : \beta_i \leq 1} \beta_i \\ &= \prod_{i \in [q+1]_0 : \beta_i \leq 1} \beta_i, \end{aligned}$$

as desired. In case 2, we have that $\alpha > 1$ and that $i = q+1$ is the only value of i for which $\beta_i \leq 1$. Hence,

$$\begin{aligned} \beta_{q+1} \min(1, \alpha) &= \beta_{q+1} \\ &= \prod_{i \in [q]_0 : \beta_i \leq 1} \beta_i, \end{aligned}$$

again as desired. This proves the induction step in the case that $\beta_{q+1} \leq 1$.

Now suppose that $\beta_{q+1} > 1$. For case 1, note that each $g_{(q+1)j}$ is 1-Hölder and recall that h is α-Hölder. Hence each $f_j = g_{(q+1)j} \circ h$ is α-Hölder. Since $\beta_{q+1} > 1$, we then have

$$\alpha = \prod_{i \in [q+1]_0 : \beta_i \leq 1} \beta_i,$$

as desired.

For case 2, distinguish two further subcases: the case that $\beta_{q+1} < \alpha$ and the case $\beta_{q+1} \geq \alpha$. We first treat the subcase $\beta_{q+1} < \alpha$. In that case, note that h and each $g_{(q+1)j}$ have all partial derivatives of order up to and including $m := \underline{\beta_{q+1}}$, and that they are Hölder continuous with exponent $\overline{\beta_{q+1}}$. Applying a multivariate version of Faà di Bruno's formula, see for example Constantine and Savits (1996), we can obtain for any $\boldsymbol{\gamma} \in \mathbb{N}_0^{d_0}$ with $|\boldsymbol{\gamma}|_0 = m$, that

$$\partial^{\gamma} f_j(\boldsymbol{x}) = \sum_{\substack{\boldsymbol{\lambda} \in \mathbb{N}_0^{d_{q+1}}: \\ 1 \leq |\boldsymbol{\lambda}| \leq m}} \sum_{i=1}^{m} \sum_{\substack{\left((\boldsymbol{v}_r)_{r=1}^{i}, (\boldsymbol{w}_r)_{r=1}^{i}\right) \\ \in S_i(\boldsymbol{\gamma}, \boldsymbol{\lambda})}} \left((\partial^{\boldsymbol{\lambda}} g_{(q+1)j}) \circ h \right) (\boldsymbol{x}) \prod_{r=1}^{i} c_{\boldsymbol{v}_r, \boldsymbol{w}_r, \boldsymbol{\gamma}} \prod_{k=1}^{d_{q+1}} \left(\partial^{\boldsymbol{v}_r} h_k(\boldsymbol{x})\right)^{w_{r,k}},$$

where for any $i \in [m]$, $S_i(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ is a finite subset with $\boldsymbol{v}_r \in \mathbb{N}_0^{d_{q+1}}$, $\boldsymbol{w}_r \in \mathbb{N}_0^{d_0}$ and $|\boldsymbol{v}_r|_0, |\boldsymbol{w}_r|_0 \leq m$ for all $r \in [i]$. Each $c_{\boldsymbol{v}_r, \boldsymbol{w}_r, \boldsymbol{\gamma}}$ is a constant depending only on $\boldsymbol{v}_r, \boldsymbol{w}_r$ and $\boldsymbol{\gamma}$. Now note that since $\partial^{\boldsymbol{\lambda}} g_{(q+1)j}$ is $\overline{\beta_{q+1}}$-Hölder and h is 1-Hölder, $\partial^{\boldsymbol{\lambda}} g_{(q+1)j} \circ h$ is also $\overline{\beta_{q+1}}$-Hölder. Similarly, $\partial^{\boldsymbol{v}_r} h_k$ is $\overline{\beta_{q+1}}$-Hölder, since $|\boldsymbol{v}_r|_0 \leq m$. Now note that products and sums of $\overline{\beta_{q+1}}$-Hölder functions are also $\overline{\beta_{q+1}}$-Hölder. This proves that $\partial^{\gamma} f_j$ is also $\overline{\beta_{q+1}}$-Hölder and hence that f_j has Hölder smoothness β_{q+1}, as desired.

For the subcase that $\beta_{q+1} \geq \alpha$, we can follow the same strategy with $|\boldsymbol{\gamma}|_0 = \underline{\alpha}$ instead of $|\boldsymbol{\gamma}|_0 = \underline{\beta_{q+1}}$. This then gives that f_j is α-Hölder, again as desired. This completes the proof of the induction step and hence of Proposition 1.

□

Proof of Proposition 2. The case $s \in \{0, 1\}$ is trivial, since then $\mathcal{F}_0(L, \boldsymbol{p}, s, F)$ contains only the zero function. Hence, suppose $s \geq 2$. Denote $|| \cdot ||_2 := || \cdot ||_{L^2([0,1]^d, \mu)}$ and $\mathcal{G} := \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$. Consider the nonparametric regression model (1) with $n \geq 3$

and μ as the distribution of $\boldsymbol{X}_1$. Let $C' > 0$ denote a generic constant depending only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F, \mu$, whose value might change from line to line. Let $\beta^{**} := \beta^*_{i^*}$ and $t^* := t_{i^*}$. Let $K' > 1$ be large enough to apply Theorem 3 to $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K')$. For this application of Theorem 3, denote the constant c in the statement of the theorem by c_3. Note that K' can be chosen depending only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, \mu$, and hence so can c. Note further that $\phi_n = n^{-(2\beta^{**})/(2\beta^{**}+t^*)}$. We will require that $K \geq K'$.

Let $\kappa > 0$ be some constant. For each sample size n, let $\hat{f}_n$ be an estimator taking values in $\mathcal{F}_0 := \mathcal{F}_0(L, \boldsymbol{p}, s, F)$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_n(\boldsymbol{X}_i))^2 - \inf_{f \in \mathcal{F}_0} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2 \leq \kappa.$$

Clearly, we have that $\Delta_n(\hat{f}_n, f_0) \leq \kappa$ for all $f_0 \in \mathcal{G}$, where $\Delta_n(\hat{f}_n, f_0)$ is as defined in (4). Then by Theorem 3 applied to $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K')$, we have

$$\sup_{f_0 \in \mathcal{G}} R(\hat{f}_n, f_0) \quad \geq \quad \sup_{f_0 \in \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K')} R(\hat{f}_n, f_0) \quad \geq \quad c_3 \phi_n \quad = \quad c_3 n^{-\frac{2\beta^{**}}{2\beta^{**}+t^*}}. \tag{36}$$

Define $\mathcal{N}_n$ as in the statement of Lemma 4 with $\mathcal{F} = \mathcal{F}_0$ and $\delta = 1/n$. Recall from the beginning of the proof of Theorem 2′ that $\mathcal{N}_n \geq 3$ if $n \geq 3$ and $s \geq 2$, so that we can apply Lemma 4′ for any $f_0 \in \mathcal{G}$ and $\epsilon \in (0, 1]$. Choosing $\epsilon = 1$ then yields

$$R(\hat{f}_n, f_0) \leq 4 \inf_{f \in \mathcal{F}_0} ||f - f_0||_2^2 + 4\kappa + C' n^{-1} \log \mathcal{N}_n.$$

Since this holds for any $f_0 \in \mathcal{G}$, it also holds if we take the supremum over all $f_0 \in \mathcal{G}$ on both the LHS and RHS. Furthermore, since this inequality holds for all $\kappa > 0$, it also holds for $\kappa = 0$. That is, we have

$$\sup_{f_0 \in \mathcal{G}} R(\hat{f}_n, f_0) \leq 4 \sup_{f_0 \in \mathcal{G}} \inf_{f \in \mathcal{F}_0} ||f - f_0||_2^2 + C' n^{-1} \log \mathcal{N}_n.$$

Now define V as in the statement of Lemma 5 and apply this lemma to the previous display to obtain

$$\sup_{f_0\in\mathcal{G}} R(\hat{f}_n, f_0) \leq 4 \sup_{f_0\in\mathcal{G}} \inf_{f\in\mathcal{F}_0} ||f - f_0||_2^2 + C'n^{-1}s\log(nLV). \tag{37}$$

Next, for any $\epsilon \in (0,1)$, define $n_\epsilon := \lfloor(\sqrt{8}\epsilon/\sqrt{c_3})^{-(2\beta^{**}+t^*)/\beta^{**}}\rfloor$. Let $c_2 \in (0,1)$ be sufficiently small such that $n_\epsilon \geq 3$ for all $\epsilon \in (0, c_2]$. Then $n_\epsilon^{-2\beta^{**}/(2\beta^{**}+t^*)} \geq 8\epsilon^2/c_3$ and $(2n_\epsilon)^{-1} \leq (\sqrt{8}\epsilon/\sqrt{c_3})^{(2\beta^{**}+t^*)/\beta^{**}}$. Letting $n = n_\epsilon$ and combining these bounds with the lower bound in (36) and the upper bound in (37), we now obtain

$$8\epsilon^2 \leq 4 \sup_{f_0\in\mathcal{G}} \inf_{f\in\mathcal{F}_0} ||f - f_0||_2^2 + C's\epsilon^{(2\beta^{**}+t^*)/\beta^{**}} \log\left(\epsilon^{-1}LV\right).$$

Now suppose $s \leq c_1\epsilon^{-t_{i^*}/\beta^*_{i^*}}/\log\left(\epsilon^{-1}LV\right)$ for some constant $c_1 > 0$. Then we can bound the upper bound of the previous display to obtain:

$$8\epsilon^2 \leq 4 \sup_{f_0\in\mathcal{G}} \inf_{f\in\mathcal{F}_0} ||f - f_0||_2^2 + C'c_1\epsilon^2.$$

We can then choose $c_1 > 0$ small enough, depending only on $q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, F, \mu$, such that the second term on the RHS can be bounded by $4\epsilon^2$. We then obtain $\epsilon \leq \sup_{f_0\in\mathcal{G}} \inf_{f\in\mathcal{F}_0} ||f - f_0||_2$, as desired.

□

Proof of Proposition 3. We first prove (i) of Proposition 3. Suppose that P is uniformly ergodic with stationary distribution π. We show that P is ϕ-irreducible with $\phi = \pi$. Indeed, suppose $A \subset \mathcal{X}$ is measurable with $\pi(A) > 0$. Then by uniform ergodicity, we have for any $z \in \mathcal{X}$ and large enough $m \in \mathbb{N}$ that $P^m(z, A) > \pi(A)/2 > 0$. Hence, P is ϕ-irreducible with $\phi = \pi$. Now suppose P is periodic. Then there exist an integer $k \geq 2$ and disjoint measurable subsets $U_1, ..., U_k \subset \mathcal{X}$ with $\pi(U_i) > 0$ for all i, such that $P(z, U_{i+1}) = 1$ for all $i \in [k-1]$ and $z \in U_i$, and such that $P(z, U_1) = 1$ for all $z \in U_k$. Let $z \in U_1$. Then we must have that $P^m(z, U_1) = 1$ if and only if m is divisible by k. By the previous argument for ϕ-irreducibility, there exists $M \in \mathbb{N}$ such that for all integers $m \geq M$, we have $P^m(z, U_1) > 0$ and $P^m(z, U_2) > 0$. Hence, $P^{kM}(z, U_1) \in (0, 1)$. However, since kM is divisible by k, we must also have $P^{kM}(z, U_1) = 1$. This is a contradiction and we conclude that P must be aperiodic. Hence, P is $\mathrm{N}(\pi)$.

Next, it follows immediately from the definition of uniform ergodicity that for $m \in \mathbb{N}$ large enough,

$$\sup_{z \in \mathcal{X}} d_{TV}(P^m(z, \cdot) - \pi) \leq 1/4.$$

That is, $t_{mix} < \infty$, with the mixing time t_{mix} as defined in Definition 1.3 of Paulin (2015). Then by Proposition 3.4 of Paulin (2015), $\gamma_{ps}(\boldsymbol{P}) \geq 1/(2t_{mix}) > 0$.

Suppose now that Q is uniformly ergodic with stationary distribution π'. We show that $P \times Q$ is also uniformly ergodic with stationary distribution $\pi \times \pi'$. Denote $R := P \times Q$. Then for any $m \in \mathbb{N}$, R^m is given by $R^m((z, z'), \cdot) = P^m(z, \cdot) \times Q^m(z', \cdot)$ for all $(z, z') \in \mathcal{X} \times \mathcal{Y}$. Note that by the definition of uniform ergodicity, there exist $M, M' > 0$ and $\rho, \psi \in (0, 1)$ such that for all $z \in \mathcal{X}$, $z' \in \mathcal{Y}$ and $m \in \mathbb{N}$, we have

$$d_{TV}\left(P^m(z, \cdot), \pi\right) \leq M\rho^m, \quad \text{and} \quad d_{TV}\left(Q^m(z', \cdot), \pi'\right) \leq M'\psi^m.$$

Now let $z \in \mathcal{X}, z' \in \mathcal{Y}$ and $m \in \mathbb{N}$. Then define the measures $\mu := P^m(z, \cdot) + \pi$ and $\mu' := Q^m(z', \cdot) + \pi'$, and the following Radon-Nikodym derivatives:

$$r := \frac{dP^m(z, \cdot)}{d\mu}, \quad r' := \frac{dQ^m(z', \cdot)}{d\mu'}, \quad p := \frac{d\pi}{d\mu}, \quad p' := \frac{d\pi'}{d\mu'}.$$

Then R has $(\mu \times \mu')$-density $r \otimes r' : (u, u') \mapsto r(u)r'(u')$ and $\pi \times \pi'$ has $(\mu \times \mu')$-density $p \otimes p' : (u, u') \mapsto p(u)p'(u')$. For any measure λ on any space $\mathcal{M}$, define the Hellinger distance between two λ-densities q, q' by $d_H(q, q') := \left[\int_{\mathcal{M}} \left(\sqrt{q} - \sqrt{q'}\right)^2 d\lambda\right]^{1/2}$. Then we can bound

$$\begin{aligned}
d_{TV}\left(R^m((z,z'),\cdot),\pi\times\pi'\right) &= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} |r\otimes r' - p\otimes p'| d(\mu\times\mu') \\
&\le d_H(r\otimes r', p\otimes p') \\
&\le \sqrt{d_H^2(r,p) + d_H^2(r',p')} \\
&\le \sqrt{\int_{\mathcal{X}} |r-p| d\mu + \int_{\mathcal{Y}} |r'-p'| d\mu'} \\
&= \sqrt{2}\cdot\sqrt{d_{TV}(P^m(z,\cdot),\,\pi) + d_{TV}(Q^m(z',\cdot),\,\pi')} \\
&\le \sqrt{2}\cdot\sqrt{M\rho^m + M'\psi^m} \\
&= \sqrt{2}\sqrt{M+M'}\left(\sqrt{\max\{\rho,\psi\}}\right)^m.
\end{aligned}$$

The first line follows from (B.1) of Ghosal and Van der Vaart (2017), the second from their Lemma B.1(i), the third from their Lemma B.8(iii), the fourth from their Lemma B.1(ii), the fifth from their (B.1) again and the penultimate line from the uniform ergodicity of P. This proves the uniform ergodicity of R, which completes the proof of (i) of Proposition 3.

We now move on to proving (ii) of Proposition 3. Denote now $R := P\times P$. Recall that for each $z, z'\in\mathcal{X}$, $R((z,z'),\cdot)$ is given by by the product measure $P(z,\cdot)\times P(z',\cdot)$. It is well-known that $L^2(\pi\times\pi)$ is isometrically isomorphic to the Hilbert tensor product $L^2(\pi)\otimes L^2(\pi)$. The isomorphism can be constructed as follows. For an orthonormal basis $\{[f_n]_\pi\}_{n\in\mathbb{N}}$ of $L^2(\pi)$, the set $\{[(z,z')\mapsto f_n(z)\cdot f_m(z')]_{\pi\times\pi}\}_{n,m\in\mathbb{N}}$ is an orthonormal basis of $L^2(\pi\times\pi)$ and we can obtain an isometric isomorphism to $L^2(\pi)\otimes L^2(\pi)$ by extending the map that sends the equivalence class $[(z,z')\mapsto f_n(z)\cdot f_m(z')]_{\pi\times\pi}$ to the tensor $[f_n]_\pi\otimes[f_m]_\pi$, for any $n,m\in\mathbb{N}$. It is then easy to see that the operator $\boldsymbol{R}$ corresponds to the operator $\boldsymbol{P}\otimes\boldsymbol{P}$ on $L^2(\pi)\otimes L^2(\pi)$, since for any $[f]_\pi,[g]_\pi\in L^2(\pi)$ and $z,z'\in\mathcal{X}$, we have that

$$\begin{aligned}\left(\boldsymbol{R}\left[(u,u')\mapsto f(u)g(u')\right]_{\pi\times\pi}\right)(z,z') &= \int_{\mathcal{X}\times\mathcal{X}} f(u)g(u')R((z,z'),d(u,u'))\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} f(u)g(u')P(z,du)P(z',du')\\ &= \left(\int_{\mathcal{X}} f(u)P(z,du)\right)\left(\int_{\mathcal{X}} g(u')P(z',du')\right)\\ &= ((\boldsymbol{P}[f]_\pi)(z))\cdot\left((\boldsymbol{P}[g]_\pi)(z')\right).\end{aligned}$$

Since P has positive pseudo spectral gap, there exists $k\in\mathbb{N}$ such that $\gamma((\boldsymbol{P}^*)^k\boldsymbol{P}^k)>0$. Denote $\boldsymbol{T}:=(\boldsymbol{P}^*)^k\,\boldsymbol{P}^k$. We use the algebraic properties of the Hilbert tensor product to calculate that

$$\begin{aligned}((\boldsymbol{P}\otimes\boldsymbol{P})^*)^k(\boldsymbol{P}\otimes\boldsymbol{P})^k &= (\boldsymbol{P}^*\otimes\boldsymbol{P}^*)^k(\boldsymbol{P}\otimes\boldsymbol{P})^k\\ &= (\boldsymbol{P}^*\otimes\boldsymbol{P}^*)^{k-1}(\boldsymbol{P}^*\boldsymbol{P}\otimes\boldsymbol{P}^*\boldsymbol{P})(\boldsymbol{P}\otimes\boldsymbol{P})^{k-1}\\ &= (\boldsymbol{P}^*\otimes\boldsymbol{P}^*)^{k-2}\left((\boldsymbol{P}^*)^2\boldsymbol{P}^2\otimes(\boldsymbol{P}^*)^2\boldsymbol{P}^2\right)(\boldsymbol{P}\otimes\boldsymbol{P})^{k-2} \qquad (38)\\ &\dots\\ &= \boldsymbol{T}\otimes\boldsymbol{T}.\end{aligned}$$

Next, for any linear operator $\boldsymbol{L}$, denote its kernel by $\mathrm{Ker}(\boldsymbol{L})$ and the dimension of this kernel by $\dim\mathrm{Ker}(\boldsymbol{L})$. Since $\gamma(\boldsymbol{T})>0$, we have $\dim\mathrm{Ker}(\boldsymbol{T}-1)=1$. We now show that $\dim\mathrm{Ker}(\boldsymbol{T}\otimes\boldsymbol{T}-1)=1$. Note first of all that $\boldsymbol{T}$ is the product of an operator and its adjoint. Hence, $\boldsymbol{T}$ is self-adjoint and positive semi-definite. Then by the Corollary to Theorem VII.3 of Reed et al. (1980), there exists a finite measure space $(M,\mathcal{M},\mu)$, a measurable function $F:M\to\mathbb{R}$ and a unitary map $\boldsymbol{U}:L^2(\pi)\to L^2(\mu)$ such that

$$\boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^{-1}=([f]_\mu\mapsto[Ff]_\mu)=:M_F.$$

Since $\boldsymbol{T}$ is positive semi-definite, F can be chosen to be non-negative. Furthermore, we have $||F||_{L^\infty(\mu)}=||T||_{L^2(\pi)}=1$. We can assume without loss of generality that $||F||_\infty=1$. Note now that

$$(\boldsymbol{U}\otimes\boldsymbol{U})(\boldsymbol{T}\otimes\boldsymbol{T})(\boldsymbol{U}^{-1}\otimes\boldsymbol{U}^{-1})=M_F\otimes M_F,$$

and that $\boldsymbol{U} \otimes \boldsymbol{U}$ is also unitary. It is then easy to see that the operator $M_F \otimes M_F$ on $L^2(\mu) \otimes L^2(\mu)$ corresponds to the operator $M_{F\times F}$ on $L^2(\mu \times \mu)$ given by

$$M_{F\times F}[f]_{\mu\times\mu} = [(x,y) \mapsto F(x)f(x,y)F(y)]_{\mu\times\mu}.$$

Hence, we have both that

$$1 = \dim \operatorname{Ker}(\boldsymbol{T} - 1) = \dim \operatorname{Ker}(M_F - 1),$$

and

$$\dim \operatorname{Ker}(\boldsymbol{T} \otimes \boldsymbol{T} - 1) = \dim \operatorname{Ker}(M_F \otimes M_F - 1) = \dim \operatorname{Ker}(M_{F\times F} - 1).$$

We show that $\dim \operatorname{Ker}(M_{F\times F} - 1) = 1$. Suppose $[f]_{\mu\times\mu} \in L^2(\mu \times \mu)$ is an eigenvector of $M_{F\times F}$ with eigenvalue 1. That is, $F(x)f(x,y)F(y) = f(x,y)$ for $(x,y) \in M \times M$ $(\mu\times\mu)$-a.e. Assume without loss of generality that this holds for all x, y. Then $F(x)F(y) = 1$ for $(x,y) \in A := f^{-1}(\mathbb{C} \setminus \{0\})$. Since $[f]_{\mu\times\mu}$ is an eigenvector, it is non-zero. That is, $(\mu \times \mu)(A) > 0$. Recall that $||F||_\infty = 1$ and that F is non-negative. Hence, we have that $F(x)F(y) = 1$ if and only if $F(x) = 1$ and $F(y) = 1$. Therefore, $A \subset F^{-1}(1) \times F^{-1}(1) =: Z \times Z$. Since A has positive measure, $Z \times Z$ has positive measure as well, and so does Z. Now note that the indicator $\mathbf{1}(Z \times Z)$ is clearly an eigenvector of $M_{F\times F}$ with eigenvalue 1. We show that f must be constant $(\mu\times\mu)$-a.e. on $Z \times Z$ and therefore that $\mathbf{1}(Z \times Z)$ spans $\operatorname{Ker}(M_{F\times F} - 1)$. Suppose it does not hold that f is constant $(\mu \times \mu)$-a.e. on $Z \times Z$. Then either its real or its imaginary part must be non-constant. Assume without loss of generality that its real part is non-constant and denote this real part by $r : M \times M \to \mathbb{R}$. Then there exists $c \in \mathbb{R}$ such that

$$(\mu \times \mu)\big((Z \times Z) \cap \{r \geq c\}\big) > 0 \quad \text{and} \quad (\mu \times \mu)\big((Z \times Z) \cap \{r < c\}\big) > 0.$$

Since $\{r \geq c\}$ and $\{r < c\}$ form a disjoint measurable partition of $M \times M$, we have that

$$(\mu \times \mu)\big((Z \times Z) \cap \{r \geq c\}\big) + (\mu \times \mu)\big((Z \times Z) \cap \{r < c\}\big) = (\mu \times \mu)(Z \times Z),$$

and hence

$$0 < (\mu \times \mu)\big((Z \times Z) \cap \{r \geq c\}\big) < (\mu \times \mu)(Z \times Z).$$

That is,

$$0 < \int_Z \int_Z \mathbf{1}\{r \geq c\} d(\mu \times \mu) < \mu(Z)\mu(Z).$$

Hence, there must exist $x \in Z$ such that

$$0 < \int_Z \mathbf{1}\{y \in Z : r(x,y) \geq c\} d\mu < \mu(Z), \quad \text{or} \quad 0 < \int_Z \mathbf{1}\{y \in Z : r(y,x) \geq c\} d\mu.$$

Suppose without loss of generality that the former holds. Since

$$\int_Z \mathbf{1}\{y \in Z : r(x,y) \geq c\} d\mu = \mu(\{y \in Z : r(x,y) \geq c\}),$$

we then have $0 < \mu(\{y \in Z : r(x,y) \geq c\}) < \mu(Z)$ and similarly that

$$\mu(\{y \in Z : r(x,y) < c\}) = \mu(Z) - \mu(\{y \in Z : r(x,y) \geq c\}) \in (0, \mu(Z)).$$

Hence, $A := \{y \in Z : r(x,y) \geq c\}$ and $B := \{y \in Z : r(x,y) < c\}$ form a disjoint measurable partition of Z and both have positive measure. Since $F = 1$ on Z, we have that the μ-equivalence classes of the two indicators $\mathbf{1}(A)$ and $\mathbf{1}(B)$ are orthogonal eigenvectors of M_F with eigenvalue 1. Since $\dim \operatorname{Ker}(M_F - 1) = 1$, this is a contradiction. Hence, $\dim \operatorname{Ker}(\boldsymbol{T} \otimes \boldsymbol{T} - 1) = 1$.

Using (38) and $\dim \operatorname{Ker}(\boldsymbol{T} \otimes \boldsymbol{T} - 1) = 1$, we have that

$$\gamma\left((\boldsymbol{R}^*)^k \boldsymbol{R}^k\right) = 1 - \sup\left(\operatorname{Spec}(\boldsymbol{T} \otimes \boldsymbol{T}) \setminus \{1\}\right). \tag{39}$$

An application of the main theorem of Brown and Pearcy (1966) yields that

$$\operatorname{Spec}(\boldsymbol{T} \otimes \boldsymbol{T}) = \{a \cdot b : a, b \in \operatorname{Spec}(\boldsymbol{T})\}. \tag{40}$$

Now recall that $\boldsymbol{T}$ is a positive semi-definite operator and hence has non-negative spectrum. Note further that $\operatorname{Spec}(\boldsymbol{T}) \setminus \{1\}$ is bounded from above by $1 - \gamma(\boldsymbol{T})$. Hence, $\operatorname{Spec}(\boldsymbol{T}) \setminus \{1\} \subset$

$[0, 1-\gamma(\boldsymbol{T})]$. Together with (40), this implies that

$$\operatorname{Spec}(\boldsymbol{T} \otimes \boldsymbol{T}) \setminus \{1\} \subset [0, (1-\gamma(\boldsymbol{T}))^2].$$

Now combine this with (39) and the inequality $(1-\gamma(\boldsymbol{T}))^2 \leq 1-\gamma(\boldsymbol{T})$ to obtain

$$\gamma\left((\boldsymbol{R}^*)^k \boldsymbol{R}^k\right) \quad \geq \quad \gamma(\boldsymbol{T}) \quad = \quad \gamma\left((\boldsymbol{P}^*)^k \boldsymbol{P}^k\right).$$

Note that this holds for any $k \in \mathbb{N}$ with $\gamma\left((\boldsymbol{P}^*)^k \boldsymbol{P}^k\right) > 0$, hence

$$\gamma_{ps}(\boldsymbol{R}) = \sup_{k \in \mathbb{N}} \frac{\gamma\left((\boldsymbol{R}^*)^k \boldsymbol{R}^k\right)}{k} \geq \sup_{k \in \mathbb{N}} \frac{\gamma\left((\boldsymbol{P}^*)^k \boldsymbol{P}^k\right)}{k} = \gamma_{ps}(\boldsymbol{P}) > 0,$$

as desired. This proves (ii) of Proposition 3.

□

Proof of Proposition 4. As pointed out in the main text, it is sufficient to prove that a modified version of Lemma 4 holds under the conditions of Proposition 4. In particular, we show that there exist constants $C_1, C_2 > 0$ depending only on d, F, ν and $\gamma_{ps}(\boldsymbol{P})$ such that for any space of measurable functions $\mathcal{F} \subset \{f : [0,1]^d \to [-F, F]\}$ with δ-covering number $\mathcal{N}_n \geq 3$, and any estimator $\hat{f}$ taking values in $\mathcal{F}$, the following holds for all $\epsilon, \delta \in (0, 1]$:

$$\begin{aligned}(1-\epsilon)^2\Delta_n - C_1\frac{\log\mathcal{N}_n}{n\epsilon} - C_1\delta &\leq R(\hat{f}, f_0) \\ &\leq (1+\epsilon)^2\left[\inf_{f \in \mathcal{F}} \mathbb{E}\left[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right] + C_2\frac{\log\mathcal{N}_n}{n\epsilon} + C_2\delta + \Delta_n\right],\end{aligned}$$

where Δ_n is as defined in the statement of Lemma 4.

We now follow the same notation that we used in the proof of Lemma 4′. Note that the following properties hold:

- $\mathbb{E}\left[\epsilon_i | (\boldsymbol{X}_j)_{j=1}^n\right] = 0$ for all i. This follows from ϵ_i being conditionally subgaussian given $(\boldsymbol{X}_j)_{j=1}^n$ and subgaussian random variables having mean zero. This also implies that its unconditional expectation is 0 and that

$$\mathbb{E}\left[\epsilon_i h\left((\boldsymbol{X}_j)_{j=1}^n\right)\right] = \mathbb{E}\left[\mathbb{E}[\epsilon_i|(\boldsymbol{X}_j)_{j=1}^n]h\left((\boldsymbol{X}_j)_{j=1}^n\right)\right] = 0,$$

for any measurable function h.

- $\mathbb{E}\left[\epsilon_i^2|(\boldsymbol{X}_j)_{j=1}^n\right] \leq \nu^2$ and hence $\mathbb{E}\left[\epsilon_i^2\right]$ for all i. This also follows from ϵ_i being conditionally ν-subgaussian given $(\boldsymbol{X}_j)_{j=1}^n$ and ν-subgaussian random variables having second moment bounded by ν^2. Both the conditional and unconditional first absolute moments $\mathbb{E}\left[|\epsilon_i|\middle|(\boldsymbol{X}_j)_{j=1}^n\right], \mathbb{E}|\epsilon_i|$ can therefore be bounded by ν, using Jensen's inequality.

These properties are already enough for most of the proof of Lemma 4′ to remain valid, mutatis mutandis, with different constants. We now point out the steps in the proof that require more involved modification. The first step that requires such involved modification is the application of Bernstein's inequality to T/F. Define $h_j(\boldsymbol{x}) := (f_j(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2$ and $\mu_j := \mathbb{E}[h_j(\boldsymbol{X})]$, for all $\boldsymbol{x} \in [0,1]^d$ and $j \in [\mathcal{N}_n]$. Then $g_j(\boldsymbol{x},\boldsymbol{y}) = h_j(\boldsymbol{y}) - h_j(\boldsymbol{x})$ for all $\boldsymbol{x},\boldsymbol{y} \in [0,1]^d$. Hence, we can bound for any $t > 0$ and any $j \in [\mathcal{N}_n]$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^n \frac{g_j(\boldsymbol{X}_i,\boldsymbol{X}_i')}{r_j F^2}\right| \geq t\right) \\
&\quad = \mathbb{P}\left(\left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i')}{r_j F^2} - n\mu_j + n\mu_j - \sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i)}{r_j F^2}\right| \geq t\right) \\
&\quad \leq \mathbb{P}\left(\left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i')}{r_j F^2} - n\mu_j\right| + \left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i)}{r_j F^2} - n\mu_j\right| \geq t\right) \\
&\quad \leq \mathbb{P}\left(\left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i')}{r_j F^2} - n\mu_j\right| \geq t/2 \text{ or } \left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i)}{r_j F^2} - n\mu_j\right| \geq t/2\right) \\
&\quad \leq \mathbb{P}\left(\left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i')}{r_j F^2} - n\mu_j\right| \geq t/2\right) + \mathbb{P}\left(\left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i)}{r_j F^2} - n\mu_j\right| \geq t/2\right) \\
&\quad = 2\mathbb{P}\left(\left|\sum_{i=1}^n \frac{h_j(\boldsymbol{X}_i)}{r_j F^2} - n\mu_j\right| \geq t/2\right),
\end{aligned} \tag{41}$$

where the first inequality follows from the triangle inequality, the second from $a < t/2$ and $b < t/2$ together implying that $a + b < t$, the third from the union bound and the final equality follows from $(\boldsymbol{X}_i)_i$ having the same distribution as $(\boldsymbol{X}_i')$. Next, arguing as we did in the proof of Lemma 4$'$ for g_j, we can now argue for h_j that the following hold for all $\boldsymbol{x} \in [0,1]^d$ and $j \in [\mathcal{N}_n]$:

$$\left| \frac{h_j(\boldsymbol{x})}{r_j F^2} - \mu_j \right| \le 4/r_j, \quad \text{and} \quad \operatorname{Var}\left(\frac{h_j(\boldsymbol{X})}{r_j F^2} \right) \le 4/F^2.$$

Since $(\boldsymbol{X}_i)_{i=1}^{\infty}$ is a strictly stationary, ϕ-irreducible and aperiodic Markov chain with unique stationary distribution π and positive pseudo spectral gap $\gamma_{ps}(\boldsymbol{P}) > 0$, we may apply Theorem 3.4 of Paulin (2015) to obtain for any $t > 0$ and any $j \in [\mathcal{N}_n]$:

$$\begin{aligned} &\mathbb{P}\left(\left| \sum_{i=1}^{n} \frac{h_j(\boldsymbol{X}_i)}{r_j F^2} \right| \ge t/2 \right) \\ &\qquad \le 2\exp\left(- \frac{t^2 \gamma_{ps}(\boldsymbol{P})}{32(n + 1/\gamma_{ps}(\boldsymbol{P})) \operatorname{Var}(h_j(\boldsymbol{X})/(r_j F^2)) + 160t/r_j} \right) \\ &\qquad \le 2\exp\left(- \frac{t^2 \gamma_{ps}(\boldsymbol{P})}{128(n + 1/\gamma_{ps}(\boldsymbol{P}))/F^2 + 160t/r_j} \right) \\ &\qquad \le 2\exp\left(- \frac{t \gamma_{ps}(\boldsymbol{P})}{256n/(t\gamma_{ps}(\boldsymbol{P})) + 160/r_j} \right). \end{aligned}$$

We now combine this with (41) and the fact that $r_j \ge \sqrt{n^{-1} \log \mathcal{N}_n}$. Hence, for $t \ge (256/\gamma_{ps}(\boldsymbol{P}))\sqrt{n \log \mathcal{N}_n}$, we have

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} \frac{g_j(\boldsymbol{X}_i, \boldsymbol{X}_i')}{r_j F^2} \right| \ge t \right) \le 4\exp\left(- \frac{t\gamma_{ps}(\boldsymbol{P})\sqrt{\log \mathcal{N}_n}}{161\sqrt{n}} \right).$$

This can then be used to bound the first and second moments of T/F as in the proof of Lemma 4$'$, but now with different constants that depend only on d, F, ν and $\gamma_{ps}(\boldsymbol{P})$. In particular, this allows us to obtain a version of (I) with different constants.

Next, we discuss the modification of the proof of (II). We can define ξ_j the same way as before, except that we now let $Z \sim N(0, \nu^2)$. Now note that conditional on $(\boldsymbol{X}_i)_{i=1}^n$, $(\epsilon_i)_{i=1}^n$ is ν-subgaussian and for each j, $\left(\frac{(f_j(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i))}{\sqrt{n}||f_j - f_0||_n} \right)_{i=1}^n$ is a vector with unit 2-norm, provided $||f_j - f_0|| \ne 0$. Hence, still conditional on $(\boldsymbol{X}_i)_{i=1}^n$, ξ_j is ν-subgaussian for each j.

We now prove the following modification of Lemma C.1: there exists a constant $C_\nu > 0$ depending only on ν such that for any $M \in \mathbb{N}$ and any real-valued ν-subgaussian random variables $\eta_1, ..., \eta_M$, we have that $\mathbb{E}\left[\max_{j\in[M]} \eta_j^2\right] \leq C_\nu(1+\log M)$. Note that it is well-known that there exists a constant $G > 0$ depending only on ν such that if Z is a real-valued ν-subgaussian random variable, then we have for every $t \geq 0$ that $\mathbb{P}(|Z| > t) \leq 2\exp(-Gt^2)$, see for example Proposition 2.5.2(i) and (v) of Vershynin (2018). We now follow the proof of Lemma C.1 with this bound instead of $\mathbb{P}(Z^2 \geq t) \leq \frac{2}{\sqrt{2\pi t}}\exp(-t/2)$. Recalling that $B := \max_{j\in[M]} \eta_j^2$, following the steps in the proof of Lemma C.1 to bound $\mathbb{E}[B]$ now yields the bound $\mathbb{E}[B] \leq T + 2M\exp(-GT)/G$ for any $T > 0$. We then choose $T = (\log M)/G$ and obtain

$$\mathbb{E}[B] \leq (\log M)/G + 2/G \leq \frac{2}{G}(1+\log M).$$

Since the constant G only depends on ν, this proves the modification of Lemma C.1. This result can then be used to bound $\mathbb{E}[\xi_{j'}^2] \leq C_\nu(1+\log M)$ in the proof of (II), in the same way as in (17). The crucial observation for this step is that conditional on $(\boldsymbol{X}_i)_{i=1}^n$, all ξ_j are ν-subgaussian, as we argued above. This bound is then sufficient to prove a version of (II) with different constants depending only on $d, F, \nu, \gamma_{ps}(\boldsymbol{P})$, by following the same steps as in the proof of Lemma 4′. The rest of the proof of Lemma 4′ remains valid, mutatis mutandis, under the conditions of Proposition 4, except that the constants might be different. This proves the modification of Lemma 4′. This modification is sufficient to prove modified versions of Theorems 2′ and 1′ with different constants. Hence, this proves Proposition 4.

□

References

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning 14*(1), 115–133.

Bauer, B. and M. Kohler (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics 47*(4), 2261–2285.

Brown, A. and C. Pearcy (1966). Spectra of tensor products of operators. *Proceedings of the American Mathematical Society 17*(1), 162–166.

Constantine, G. and T. Savits (1996). A multivariate Faa di Bruno formula with applications. *Transactions of the American Mathematical Society 348*(2), 503–520.

Douc, R., E. Moulines, P. Priouret, and P. Soulier (2018). *Markov chains.* Springer.

Evci, U., T. Gale, J. Menick, P. S. Castro, and E. Elsen (2020). Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, pp. 2943–2952. PMLR.

Fan, J., B. Jiang, and Q. Sun (2021). Hoeffding's inequality for general Markov chains and its applications to statistical learning. *J. Mach. Learn. Res. 22*, 139–1.

Fan, J., C. Ma, and Y. Zhong (2021). A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics 36*(2), 264.

Ghosal, S. and A. Van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, Volume 44. Cambridge University Press.

Grenander, U. and G. Szegö (1958). *Toeplitz forms and their applications.* University of California Press.

Gunasekar, S., J. Lee, D. Soudry, and N. Srebro (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(4), 757–779.

He, F., B. Wang, and D. Tao (2019). Piecewise linear activations substantially shape the loss surfaces of neural networks. In *International Conference on Learning Representations*.

Iizuka, S., E. Simo-Serra, and H. Ishikawa (2016). Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG) 35*(4), 1–11.

Jiang, B., Q. Sun, and J. Fan (2020). Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721v2*.

Kallenberg, O. (1997). *Foundations of modern probability*, Volume 2. Springer.

Le Cun, Y., J. Denker, and S. Solla (1989). Optimal brain damage. *Advances in neural information processing systems 2*.

Liu, B., Z. Liu, T. Zhang, and T. Yuan (2021). Non-differentiable saddle points and sub-optimal local minima exist for deep ReLU networks. *Neural Networks 144*, 75–89.

Lv, Y., Y. Duan, W. Kang, Z. Li, and F.-Y. Wang (2014). Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems 16*(2), 865–873.

Meyn, S. P. and R. L. Tweedie (2012). *Markov chains and stochastic stability.* Springer Science & Business Media.

Park, B. U., E. Mammen, Y. K. Lee, and E. R. Lee (2015). Varying coefficient regression models: a review and new developments. *International Statistical Review 83*(1), 36–64.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability 20*, 1–32.

Reed, M., B. Simon, et al. (1980). *Methods of Modern Mathematical Physics I: Functional analysis*, Volume 1. Gulf Professional Publishing.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics 48*(4), 1875–1897.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), 1040–1053.

Thakkar, A. and K. Chaudhari (2021). A comprehensive survey on deep neural networks for stock market: the need, challenges, and future directions. *Expert Systems with Applications 177*, 114800.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer New York, NY.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Wong, H., R. Zhang, W.-c. Ip, and G. Li (2008). Functional-coefficient partially linear regression model. *Journal of Multivariate Analysis 99*(2), 278–305.

Yang, L. and A. Shami (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing 415*, 295–316.