Covering Numbers for Deep ReLU Networks with Applications to Function Approximation and Nonparametric Regression

Weigutian Ou
ETH Zürich
wou@mins.ee.ethz.ch

Helmut Bölcskei
ETH Zürich
hboelcskei@ethz.ch

Abstract. Covering numbers of families of (deep) ReLU networks have been used to characterize their approximation-theoretic performance, upper-bound the prediction error they incur in nonparametric regression, and quantify their classification capacity. These results are based on covering number upper bounds obtained through the explicit construction of coverings. Lower bounds on covering numbers do not seem to be available in the literature. The present paper fills this gap by deriving tight (up to a multiplicative constant) lower and upper bounds on the covering numbers of fully-connected networks with bounded weights, sparse networks with bounded weights, and fully-connected networks with quantized weights. Thanks to the tightness of the bounds, a fundamental understanding of the impact of sparsity, quantization, bounded vs. unbounded weights, and network output truncation can be developed. Furthermore, the bounds allow to characterize the fundamental limits of neural network transformation, including network compression, and lead to sharp upper bounds on the prediction error in nonparametric regression through deep networks. Specifically, we can remove a $\log^6(n)$-factor in the best known sample complexity rate in the estimation of Lipschitz functions through deep networks thereby establishing optimality. Finally, we identify a systematic relation between optimal nonparametric regression and optimal approximation through deep networks, unifying numerous results in the literature and uncovering general underlying principles.

Keywords: Deep neural networks, machine learning, covering number, approximation theory, nonparametric regression

Mathematics Subject Classification: 68T07, 41A25, 62G08

1 Introduction

It is well known that neural networks exhibit universal approximation properties [1, 2, 3, 4], but these results typically require infinitely large, specifically infinitely wide, networks. Neural networks employed in practice are, however, subject to constraints on width, depth,

weight magnitude and precision, and connectivity (i.e., the number of nonzero weights). To characterize the performance limits of neural networks under such constraints, it is necessary to quantify the complexity of the function classes they realize. This is typically done through two widely used complexity notions, namely Vapnik-Chervonenkis (VC) dimension [5] and covering numbers [6].

The VC dimension finds application in the characterization of (i) the approximation-theoretic limits of neural networks with the ReLU activation function, see e.g. [7], hereafter referred to as ReLU networks, and (ii) the prediction error incurred in nonparametric regression through ReLU networks, see e.g. [8]. Nearly-tight bounds on the VC dimension of ReLU networks were reported in [9], specifically upper and lower bounds differing only by a multiplicative factor of order lower than that of the upper and the lower bound.

Covering numbers have been used to characterize the approximation-theoretic limits of ReLU networks [10, 11, 12], upper-bound the prediction error they incur in nonparametric regression [13, 14], and quantify their classification capacity [15, 16, 17]. These analyses typically construct coverings by quantizing the network weights to a precision commensurate with the desired covering ball radius. The cardinality of the resulting coverings then provides upper bounds on the covering number. Corresponding explicit lower bounds are, to the best of our knowledge, not available in the literature.

The contributions of the present paper can be organized along three main threads. The first one revolves around explicit lower bounds on the covering number of fully-connected ReLU networks with uniformly bounded weights. In particular, these bounds are shown, by way of establishing matching upper bounds, to be tight in terms of scaling behavior. The techniques we devise to derive the bounds are novel and partly rely on results recently reported by the authors of the present paper in [18].

The second thread of contributions illustrates, by way of application scenarios, what is made possible by the tightness of the covering number bounds identified before. The first scenario is concerned with the fundamental limits of neural network transformation, concretely, the approximation of a given class of networks by another class that is subject to different constraints. This includes applications such as network quantization [19] and network compression [20]. In the second scenario, we consider the fundamental limits of function approximation through ReLU networks. A novel minimax error upper bound, interesting in its own right, is shown to lead to sharp upper bounds on the prediction error in nonparametric regression through ReLU networks. This result also allows us to uncover a systematic relation between optimal nonparametric regression and optimal approximation through (deep) ReLU networks thereby unifying numerous corresponding results in the literature [13, 14, 21] and identifying general underlying principles. In all cases considered, we either improve upon best known results in the literature or fill gaps in available theories.

Our third objective is to establish tight covering number bounds for sparse (in terms of connectivity) networks with bounded weights and for fully-connected networks with quantized weights. We also provide an upper bound on the covering number of fully-connected networks with unbounded weights and truncated outputs. These three choices are motivated by their prevalence in theoretical analyses and practical applications, see e.g. [10, 13, 14, 20, 21, 22, 23, 24, 25].

The remainder of the paper is organized as follows. Frequently used definitions are provided at the end of this section, while basic notation and further definitions are listed in Appendix A. In Section 2, we present our results on the covering number of fully-connected ReLU networks with uniformly bounded weights. Sections 3 and 4 discuss the application of our covering number bounds to neural network transformation, function approximation, and nonparametric regression. Sections 5-7 report the covering number bounds for sparse networks with uniformly bounded weights, fully-connected networks with quantized weights, and fully-connected

networks with truncated outputs, respectively.

1.1 Important Definitions

We start with the definition of ReLU networks.

Definition 1.1. *Let $L, N_0, N_1, \ldots, N_L \in \mathbb{N}$. A network configuration Φ is a sequence of matrix-vector tuples*

$$\Phi = ((A_i, b_i))_{i=1}^L,$$

with $A_i \in \mathbb{R}^{N_i \times N_{i-1}}$, $b_i \in \mathbb{R}^{N_i}$, $i = 1, \ldots, L$. We refer to N_i as the width of the i-th layer, $i = 0, \ldots, L$, and call the tuple $(N_0, \ldots, N_L)$ the architecture of the network configuration. $\mathcal{N}(d)$ denotes the set of all network configurations with input dimension $N_0 = d$ and output dimension $N_L = 1$. The depth of the configuration Φ is $\mathcal{L}(\Phi) := L$, its width $\mathcal{W}(\Phi) := \max_{i=0,\ldots,L} N_i$, its weight set $\mathrm{coef}(\Phi) := \bigcup_{i=1,\ldots,L}(\mathrm{coef}(A_i) \bigcup \mathrm{coef}(b_i))$, with $\mathrm{coef}(A)$ and $\mathrm{coef}(b)$ denoting the value set of the entries of A and b, respectively, its weight magnitude $\mathcal{B}(\Phi) := \max_{i=1,\ldots,L} \max\{\|A_i\|_\infty, \|b_i\|_\infty\}$, and its connectivity $\mathcal{M}(\Phi) := \sum_{\ell=1}^L (\|A_\ell\|_0 + \|b_\ell\|_0)$.

We define, recursively, the network realization $R(\Phi) : \mathbb{R}^{N_0} \mapsto \mathbb{R}^{N_L}$, associated with the network configuration Φ, according to

$$R(\Phi) = \begin{cases} S(A_L, b_L), & \text{if } L = 1, \\ S(A_L, b_L) \circ \rho \circ R(((A_i, b_i))_{i=1}^{L-1}), & \text{if } L \geq 2, \end{cases} \tag{1}$$

where $S(A, b)$ is the affine mapping $S(A, b)(x) = Ax + b, x \in \mathbb{R}^{n_2}$, with $A \in \mathbb{R}^{n_1 \times n_2}$, $b \in \mathbb{R}^{n_1}$, and $\rho(x) := \max\{x, 0\}$, for $x \in \mathbb{R}$, is the ReLU activation function, which, when applied to vectors, acts elementwise.

The family of network configurations with depth at most L, width at most W, weight magnitude at most B, where $B \in \mathbb{R}_+ \cup \{\infty\}$, connectivity at most s, weights taking values in $\mathbb{A} \subseteq \mathbb{R}$, d-dimensional input, and 1-dimensional output, for $d \in \mathbb{N}$, $W, L, s \in \mathbb{N} \cup \{\infty\}$, with[1] $W \geq d$, is denoted as

$$\mathcal{N}_{\mathbb{A}}(d, W, L, B, s) = \{\Phi \in \mathcal{N}(d) : \mathcal{W}(\Phi) \leq W, \ \mathcal{L}(\Phi) \leq L, \ \mathcal{B}(\Phi) \leq B, \mathcal{M}(\Phi) \leq s, \mathrm{coef}(\Phi) \subseteq \mathbb{A}\},$$

with the family of associated network realizations

$$\mathcal{R}_{\mathbb{A}}(d, W, L, B, s) := \{R(\Phi) : \Phi \in \mathcal{N}_{\mathbb{A}}(d, W, L, B, s)\}. \tag{2}$$

To simplify notation, for $\mathbb{A} = \mathbb{R}$, we allow omission of the argument $\mathbb{A}$ in $\mathcal{N}_{\mathbb{A}}(d, W, L, B, s)$ and $\mathcal{R}_{\mathbb{A}}(d, W, L, B, s)$. When $s = \infty$, we allow omission of the argument s in $\mathcal{N}_{\mathbb{A}}(d, W, L, B, s)$ and $\mathcal{R}_{\mathbb{A}}(d, W, L, B, s)$. Furthermore, we allow omission of both arguments B, s in $\mathcal{N}_{\mathbb{A}}(d, W, L, B, s)$ and $\mathcal{R}_{\mathbb{A}}(d, W, L, B, s)$ when $B = s = \infty$. One specific, frequently used, incarnation of these policies is $\mathcal{N}(d, W, L) = \mathcal{N}_{\mathbb{R}}(d, W, L, \infty, \infty)$ and $\mathcal{R}(d, W, L) = \mathcal{R}_{\mathbb{R}}(d, W, L, \infty, \infty)$.

We emphasize the importance of differentiating between network configurations and network realizations. Different network configurations may result in the same realization. Nevertheless, whenever there is no potential for confusion, we shall use the term network to collectively refer to both configurations and realizations.

Throughout the paper, we shall frequently use the covering number and the packing number, defined as follows.

---

[1] The condition $W \geq d$ is formally stated here so as to prevent the trivial case of $\mathcal{N}_{\mathbb{A}}(d, W, L, B, s)$ being an empty set. It will be a standing assumption throughout the paper.

Definition 1.2 (Covering number and packing number). *[6, Definitions 5.1 and 5.4] Let $(\mathcal{Y}, \delta)$ be a metric space. An ε-covering of $\mathcal{X} \subseteq \mathcal{Y}$ is a subset $\{x_1, \ldots, x_n\}$ of $\mathcal{X}$ such that for all $x \in \mathcal{X}$, there exists an $i \in \{1, \ldots, n\}$ so that $\delta(x, x_i) \leq \varepsilon$. The ε-covering number $N(\varepsilon, \mathcal{X}, \delta)$ is the cardinality of a smallest ε-covering of $\mathcal{X}$. An ε-packing of $\mathcal{X}$ is a subset $\{x_1, \ldots, x_n\}$ of $\mathcal{X}$ such that $\delta(x_i, x_j) > \varepsilon$, for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$. The ε-packing number $M(\varepsilon, \mathcal{X}, \delta)$ is the cardinality of a largest ε-packing of $\mathcal{X}$.*

To simplify notation, when δ is the $L^p(\mathbb{X})$-norm, with $\mathbb{X} \subseteq \mathbb{R}^d$ and $p \in [1, \infty]$, we may write $N(\varepsilon, \mathcal{F}, L^p(\mathbb{X})) := N(\varepsilon, \mathcal{F}, \|\cdot\|_{L^p(\mathbb{X})})$. Moreover, we shall use $N(\varepsilon, \mathcal{F}, L^2(P)) := N(\varepsilon, \mathcal{F}, \|\cdot\|_{L^2(P)})$, for P a distribution on $\mathbb{X}$. The same conventions apply to the packing number.

2 Fully-connected ReLU Networks with Uniformly Bounded Weights

Our covering number bounds for fully-connected ReLU networks with uniformly bounded weights are as follows.

Theorem 2.1. *Let $p \in [1, \infty]$, $d, W, L \in \mathbb{N}$, $B, \varepsilon \in \mathbb{R}_+$ with $B \geq 1$ and $\varepsilon \in (0, 1/2)$. We have*

$$\log(N(\varepsilon, \mathcal{R}(d, W, L, B), L^p([0,1]^d))) \leq C W^2 L \log\left(\frac{(W+1)^L B^L}{\varepsilon}\right), \tag{3}$$

where $C \in \mathbb{R}_+$ is an absolute constant. Further, if, in addition, $W, L \geq 60$, then

$$\log(N(\varepsilon, \mathcal{R}(d, W, L, B), L^p([0,1]^d))) \geq c\, W^2 L \log\left(\frac{(W+1)^L B^L}{\varepsilon}\right), \tag{4}$$

where $c \in \mathbb{R}_+$ is an absolute constant.

Proof. The proofs of the upper bound and the lower bound are provided in Sections 2.1 and 2.2, respectively. $\square$

We remark that, for $W, L \geq 60$ and $\varepsilon \in (0, 1/2)$, the upper bound (3) and the lower bound (4) differ only by the multiplicative absolute constants C, c to be specified in the proof. These constants as well as the condition $W, L \geq 60$ are chosen for expositional convenience of the proof; improvements are possible, but will not be pursued here.

The covering number upper bounds available in the literature apply to specific settings. For example, [13, Lemma 5] addresses the case $B = 1$ and [14, Lemma 5.3] pertains to $p = \infty$, while we consider the general case $B \in [1, \infty)$, $p \in [1, \infty]$. We note that [13, Lemma 5] and [14, Lemma 5.3] also apply to sparse ReLU networks with uniformly bounded weights as considered in Section 5, again for $B \in [1, \infty)$, $p \in [1, \infty]$.

2.1 Proof of the Upper Bound in Theorem 2.1

The proof is effected by constructing an explicit ε-covering of $\mathcal{R}(d, W, L, B)$ with elements in $\mathcal{R}_{[-B,B] \cap 2^{-b}\mathbb{Z}}(d, W, L)$, where $b \in \mathbb{N}$ is a parameter suitably depending on ε. We start with three technical lemmata, and then provide the proof of the upper bound at the end of the section. The first lemma quantifies the distance between the realizations of two networks sharing the same architecture.

Lemma 2.2. *[18, Lemma E.1] Let $d, W, L, \ell \in \mathbb{N}$ with $\ell \leq L$, $B \in \mathbb{R}_+$ with $B \geq 1$, and let*

$$\Phi^i = ((A_j^i, b_j^i))_{j=1}^{\ell} \in \mathcal{N}(d, W, L, B), \quad i = 1, 2,$$

have the same architecture. Then,

$$\|R(\Phi^1) - R(\Phi^2)\|_{L^{\infty}([0,1]^d)} \leq L(W+1)^L B^{L-1} \|\Phi^1 - \Phi^2\|, \tag{5}$$

where

$$\|\Phi^1 - \Phi^2\| := \max_{j=1,\dots,\ell} \max\{\|A_j^1 - A_j^2\|_{\infty}, \|b_j^1 - b_j^2\|_{\infty}\}. \tag{6}$$

Based on Lemma 2.2, we now construct the announced ε-covering of $\mathcal{R}(d, W, L, B)$.

Lemma 2.3. *Let $p \in [1, \infty]$, $d, W, L, b \in \mathbb{N}$, and $B \in \mathbb{R}_+$ with $B \geq 1$. Then, the set $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d, W, L)$ is an $(L(W+1)^L B^{L-1} 2^{-b})$-covering of $\mathcal{R}(d, W, L, B)$ with respect to the $L^p([0,1]^d)$-norm.*

Proof. Define the quantization mapping $q_b : [-B, B] \mapsto [-B, B] \cap 2^{-b}\mathbb{Z}$ as

$$q_b(x) = \begin{cases} 2^{-b}\lfloor 2^b x \rfloor, & \text{for } x \in [0, B], \\ 2^{-b}\lceil 2^b x \rceil, & \text{for } x \in [-B, 0), \end{cases}$$

and note that $|x - q_b(x)| \leq 2^{-b}$, for all $x \in [-B, B]$. When applied to matrices or vectors, $q_b(\cdot)$ acts elementwise. Now, arbitrarily fix $R(\Phi) \in \mathcal{R}(d, W, L, B)$ with $\Phi = ((A_\ell, b_\ell))_{\ell=1}^{\tilde{L}} \in \mathcal{N}(d, W, L, B)$ and $\tilde{L} \leq L$, and quantize the weights of Φ according to

$$Q_b(\Phi) = ((q_b(A_\ell), q_b(b_\ell)))_{\ell=1}^{\tilde{L}} \in \mathcal{N}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d, W, L, B).$$

We then have

$$\|\Phi - Q_b(\Phi)\| = \max_{\ell=1,\dots,\tilde{L}} \max\{\|A_\ell - q_b(A_\ell)\|_{\infty}, \|b_\ell - q_b(b_\ell)\|_{\infty}\} \leq 2^{-b},$$

which, together with Lemma 2.2, yields

$$\|R(\Phi) - R(Q_b(\Phi))\|_{L^{\infty}([0,1]^d)} \leq L(W+1)^L B^{L-1} 2^{-b}. \tag{7}$$

As

$$\|R(\Phi) - R(Q_b(\Phi))\|_{L^p([0,1]^d)} \leq \sup_{x \in [0,1]^d} |R(\Phi)(x) - R(Q_b(\Phi))(x)| = \|R(\Phi) - R(Q_b(\Phi))\|_{L^{\infty}([0,1]^d)},$$

(7) implies that

$$\|R(\Phi) - R(Q_b(\Phi))\|_{L^p([0,1]^d)} \leq L(W+1)^L B^{L-1} 2^{-b}. \tag{8}$$

We can therefore conclude that $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d, W, L)$ is an $(L(W+1)^L B^{L-1} 2^{-b})$-covering of $\mathcal{R}(d, W, L, B)$ in the $L^p([0,1]^d)$-norm. □

It remains to upper-bound the cardinality of the covering $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d, W, L)$ identified in Lemma 2.3. To this end, we first state an auxiliary result from [18].

Lemma 2.4. *[18, Proposition 2.4] For $d, W, L \in \mathbb{N}$ and a finite set $\mathbb{A} \subseteq \mathbb{R}$ with $|\mathbb{A}| \geq 2$, it holds that*

$$\log(|\mathcal{R}_{\mathbb{A}}(d, W, L)|) \leq \log(|\mathcal{N}_{\mathbb{A}}(d, W, L)|) \leq 5W^2 L \log(|\mathbb{A}|). \tag{9}$$

We next make the choice of b explicit. Specifically, we set

$$b := \left\lceil \log\left(\frac{L(W+1)^L B^{L-1}}{\varepsilon}\right) \right\rceil. \tag{10}$$

Noting that $L(W+1)^L B^{L-1} 2^{-b} \leq \varepsilon$, it follows from Lemma 2.3 that $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L)$ is an ε-covering of $\mathcal{R}(d,W,L,B)$ with respect to the $L^p([0,1]^d)$-norm. By minimality of the covering number, we have

$$N(\varepsilon, \mathcal{R}(d,W,L,B), L^p([0,1]^d)) \leq |\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L)|. \tag{11}$$

Application of Lemma 2.4 yields an upper bound on the cardinality of the covering according to

$$\log(|\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L)|) \leq 5W^2 L \log(|[-B,B]\cap 2^{-b}\mathbb{Z}|). \tag{12}$$

The term $\log(|[-B,B]\cap 2^{-b}\mathbb{Z}|)$ can now be bounded as follows

$$\log(|[-B,B]\cap 2^{-b}\mathbb{Z}|) = \log(|[-2^b B, 2^b B]\cap \mathbb{Z}|) \tag{13}$$
$$\leq \log(\lfloor 2\cdot 2^b B + 1\rfloor) \tag{14}$$
$$\leq \log(4\cdot 2^b B) \tag{15}$$
$$= 2 + \left\lceil \log\left(\frac{L(W+1)^L B^{L-1}}{\varepsilon}\right) \right\rceil + \log(B) \tag{16}$$
$$\leq 3 + \log\left(\frac{L(W+1)^L B^{L-1}}{\varepsilon}\right) + \log(B) \tag{17}$$
$$\leq 3\log\left(\frac{L(W+1)^L B^L}{\varepsilon}\right), \tag{18}$$

where (15) is by $1 \leq 2\cdot 2^b B$ and in (18) we used $3 \leq 2\log\left(\frac{L(W+1)^L B^L}{\varepsilon}\right)$ owing to $\varepsilon \in (0,1/2)$. Putting (11)-(18) together, yields

$$\log(N(\varepsilon, \mathcal{R}(d,W,L,B), L^p([0,1]^d))) \leq 15W^2 L \log\left(\frac{L(W+1)^L B^L}{\varepsilon}\right) \tag{19}$$
$$\leq 30W^2 L \log\left(\frac{(W+1)^L B^L}{\varepsilon}\right) \tag{20}$$

where (20) follows from $\frac{L(W+1)^L B^L}{\varepsilon} \leq \frac{(W+1)^L\cdot(W+1)^L B^L}{\varepsilon} \leq \left(\frac{(W+1)^L B^L}{\varepsilon}\right)^2$. The proof is concluded by taking $C := 30$.

2.2 Proof of the Lower Bound in Theorem 2.1

We again start with a series of technical results, which will then be synthesized to the proof of the lower bound. The first of these results reduces the problem of lower-bounding the covering number of $\mathcal{R}(d,W,L,B)$ with respect to the $L^p([0,1]^d)$-norm to that of lower-bounding the packing number of $\mathcal{R}(1,W,L,B)$ with respect to the $L^1([0,1])$-norm.

Lemma 2.5. *Let $p \in [1,\infty]$, $d,W,L \in \mathbb{N}$, $B,\varepsilon \in \mathbb{R}_+$ with $B \geq 1$ and $\varepsilon \in (0,1/2)$. We have*

$$N(\varepsilon, \mathcal{R}(d,W,L,B), L^p([0,1]^d)) \geq M(2\varepsilon, \mathcal{R}(d,W,L,B), L^p([0,1]^d))) \tag{21}$$
$$\geq M(2\varepsilon, \mathcal{R}(1,W,L,B), L^1([0,1]))). \tag{22}$$

Proof. The inequality (21) follows from Lemma F.1. To establish (22), we show that a maximal (2ε)-packing of $\mathcal{R}(1, W, L, B)$ with respect to the $L^1([0,1])$-norm induces a (2ε)-packing of $\mathcal{R}(d, W, L, B)$ with respect to the $L^p([0,1]^d)$-norm and of the same cardinality. The proof of this statement is provided in Appendix B.2. $\square$

We shall next make use of the fact that ReLU networks can efficiently realize one-dimensional bounded continuous piecewise linear functions, defined as follows.

Definition 2.6 (One-dimensional bounded continuous piecewise linear functions)**.** *[18, Definition B.2] Let $M \in \mathbb{N}$, with $M \geq 3$, $E \in \mathbb{R}_+ \cup \{\infty\}$, and let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $\mathbb{R}$. Define the set of functions*

$$\Sigma(X, E) = \big\{ f \in C(\mathbb{R}) : \|f\|_{L^\infty(\mathbb{R})} \leq E, f \text{ is constant on } (-\infty, x_0] \text{ and } [x_{M-1}, \infty),$$
$$f \text{ is affine on } [x_i, x_{i+1}], i = 0, \ldots, M - 2 \big\}.$$

For a function $f \in \Sigma(X, E)$, we call X the set of its breakpoints, as the slope of f can change only at these points. We refer to the intervals $(-\infty, x_0], [x_i, x_{i+1}], i = 0, \ldots, M - 2, [x_{M-1}, \infty)$ as the piecewise linear regions of f.

We only need to consider breakpoint sets of the form

$$X_N := (i/N)_{i=0}^N, \quad N \in \mathbb{N},$$

along with the associated function families $\Sigma(X_N, E)$, $E \in \mathbb{R}_+$, whose $L^1([0,1])$-covering number can be lower-bounded as follows.

Lemma 2.7. *For $N \in \mathbb{N}$, $\varepsilon, E \in \mathbb{R}_+$, we have*

$$\log(M(\varepsilon, \Sigma(X_N, E), L^1([0,1]))) \geq N \log\left(\left\lceil \frac{E}{4\varepsilon N} \right\rceil\right), \tag{23}$$

with $X_N = (i/N)_{i=0}^N$.

Proof. See Appendix B.3. $\square$

To realize functions in $\Sigma(X_N, E)$ efficiently by ReLU networks, we need two technical results from [18], which we restate for convenience.

Lemma 2.8. *[18, Proposition C.1] Let $M \in \mathbb{N}$ with $M \geq 3$, $E \in \mathbb{R}_+$, and let $X = (x_i)_{i=0}^{M-1}$ be a strictly increasing sequence taking values in $[0,1]$. Then, for all $u, v \in \mathbb{N}$ such that $u^2 v \geq M$, we have*

$$\Sigma(X, E) \subseteq \mathcal{R}(1, 20u, 30v, \max\{1, C_k M^6 (R_m(X))^4 E\}),$$

for an absolute constant $C_k \in \mathbb{R}$ satisfying $2 \leq C_k \leq 10^5$, and where $R_m(X) := \max_{i=1,\ldots,M}(x_i - x_{i-1})^{-1}$.

The second result is as follows.

Lemma 2.9. *[18, Proposition H.4] Let $W_1, L_1 \in \mathbb{N}$, with $W_1 \geq 2$, $L_2 \in \mathbb{N} \cup \{0\}$, and $B_1, B_2 \in \mathbb{R}_+$, with $B_1, B_2 \geq 1$. It holds that*

$$\frac{(B_2)^{L_1+L_2} \lfloor W_1/2 \rfloor^{L_2}}{B_1^{L_1}} \cdot \mathcal{R}(1, W_1, L_1, B_1) \subseteq \mathcal{R}(1, W_1, L_1 + L_2, B_2). \tag{24}$$

We are now ready to prove the lower bound in Theorem 2.1 and start by noting that thanks to Lemma 2.5, it suffices to lower-bound $M(2\varepsilon, \mathcal{R}(1, W, L, B), L^1([0,1]))$. We proceed to identify the family of bounded continuous piecewise linear functions corresponding to the set $\mathcal{R}(1, W, L, B)$. To this end, we first introduce notation, namely we set

$$u := \left\lfloor \frac{W}{20} \right\rfloor, \quad v := \left\lfloor \frac{L}{60} \right\rfloor.$$

As $W, L \geq 60$, we have $u \geq 3$ and $v \geq 1$. Application of Lemma 2.8 with $M = u^2 v$, $X = X_{u^2 v - 1} = (\frac{i}{u^2 v - 1})_{i=0}^{u^2 v - 1}$, $R_m(X_{u^2 v - 1}) = u^2 v - 1$, and $E = \frac{1}{C_k (u^2 v)^{10}}$, with the absolute constant C_k per Lemma 2.8, yields

$$\Sigma\left(X_{u^2 v - 1}, \frac{1}{C_k (u^2 v)^{10}}\right) \subseteq \mathcal{R}(1, 20u, 30v, 1). \tag{25}$$

Next, application of Lemma 2.9 with $W_1 = 20u \geq 2$, $L_1 = 30v$, $B_1 = 1$, $L_2 = 30v$, $B_2 = B \geq 1$, yields $(B^{60v}(10u)^{30v}) \cdot \mathcal{R}(1, 20u, 30v, 1) \subseteq \mathcal{R}(1, 20u, 60v, B)$, which together with $\mathcal{R}(1, 20u, 60v, B) = \mathcal{R}(1, 20\lfloor\frac{W}{20}\rfloor, 60\lfloor\frac{L}{60}\rfloor, B) \subseteq \mathcal{R}(1, W, L, B)$ establishes that

$$(B^{60v}(10u)^{30v}) \cdot \mathcal{R}(1, 20u, 30v, 1) \subseteq \mathcal{R}(1, W, L, B). \tag{26}$$

Moreover, as $a \cdot \Sigma(X_{u^2 v - 1}, b) = \Sigma(X_{u^2 v - 1}, ab)$, for all $a, b \in \mathbb{R}_+$, we have

$$(B^{60v}(10u)^{30v}) \cdot \Sigma\left(X_{u^2 v - 1}, \frac{1}{C_k (u^2 v)^{10}}\right) = \Sigma\left(X_{u^2 v - 1}, \frac{(B^{60v}(10u)^{30v})}{C_k (u^2 v)^{10}}\right). \tag{27}$$

Combining (26), (27), and (25), then yields

$$\Sigma\left(X_{u^2 v - 1}, \frac{(B^{60v}(10u)^{30v})}{C_k (u^2 v)^{10}}\right) \subseteq \mathcal{R}(1, W, L, B). \tag{28}$$

We have

$$\log(M(2\varepsilon, \mathcal{R}(1, W, L, B), L^1([0,1]))) \tag{29}$$

$$\geq \log\left(M\left(2\varepsilon, \Sigma\left(X_{u^2 v - 1}, \frac{B^{60v}(10u)^{30v}}{C_k (u^2 v)^{10}}\right), L^1([0,1])\right)\right) \tag{30}$$

$$\geq (u^2 v - 1) \log\left(\left\lceil \frac{B^{60v}(10u)^{30v}}{C_k (u^2 v)^{10}} \cdot \frac{1}{4 \cdot 2\varepsilon(u^2 v - 1)} \right\rceil\right) \tag{31}$$

$$\geq (u^2 v - 1) \log\left(\frac{B^{60v}(100u^2)^v}{\varepsilon} \cdot \frac{10^6}{8C_k}\right) \tag{32}$$

$$\geq (u^2 v - 1) \log\left(\frac{(100Bu)^v}{\varepsilon}\right) \tag{33}$$

where in (30) we used the inclusion relation (28) together with Lemma F.1, (31) is by Lemma 2.7, (32) follows from $(10u)^{30v} \geq 10^6 \cdot u^{22} \cdot (2^v)^{11} \cdot (10u)^{2v} \geq 10^6 \cdot u^{22} \cdot v^{11} \cdot (10u)^{2v} \geq 10^6(u^2 v)^{10}(u^2 v - 1)(100u^2)^v$, and in (33) we employed $u \geq 3$, $v \geq 1$, $B \geq 1$, and $8C_k \leq 8 \cdot 10^5 < 10^6$. We next

lower-bound $(u^2v - 1)\log(\frac{(100Bu)^v}{\varepsilon})$ in terms of W, L, B according to

$$(u^2v - 1)\log\left(\frac{(100Bu)^v}{\varepsilon}\right) \geq \frac{1}{2}u^2v\log\left(\frac{(100Bu)^v}{\varepsilon}\right) \tag{34}$$

$$\geq \frac{1}{2}\left(\frac{W}{40}\right)^2\frac{L}{120}\log\left(\frac{(100B\frac{W}{40})^{\frac{L}{120}}}{\varepsilon}\right) \tag{35}$$

$$\geq \frac{1}{2\cdot40^2\cdot120}W^2L\log\left(\frac{(BW)^{\frac{L}{120}}}{\varepsilon^{\frac{1}{120}}}\right) \tag{36}$$

$$= \frac{1}{2\cdot40^2\cdot120^2}W^2L\log\left(\frac{W^LB^L}{\varepsilon}\right). \tag{37}$$

$$\geq \frac{1}{4\cdot40^2\cdot120^2}W^2L\log\left(\frac{(W+1)^LB^L}{\varepsilon}\right), \tag{38}$$

where (34) follows from $u^2v - 1 \geq u^2v - \frac{1}{9}u^2v \geq \frac{1}{2}u^2v$ as $u \geq 3$ and $v \geq 1$, in (35) we used $u = \lfloor\frac{W}{20}\rfloor \geq \frac{W}{40}$ and $v = \lfloor\frac{L}{60}\rfloor \geq \frac{L}{120}$, and (38) is by $2\log(\frac{W^LB^L}{\varepsilon}) = \log(\frac{W^{2L}B^{2L}}{\varepsilon^2}) \geq \log(\frac{(W+1)^LB^L}{\varepsilon})$ as $W \geq 60$ by assumption. The proof is concluded by setting $c = \frac{1}{4\cdot40^2\cdot120^2}$.

3 Neural Network Transformation and Function Approximation

We now show how the precise characterization of ReLU network covering numbers obtained in the previous section can be put to work to characterize the fundamental limits of neural network transformation and function approximation. Before describing the specifics of these two problems, we need a general result which relates the covering numbers of sets $\mathcal{G}$ and $\mathcal{F}$ that are close in terms of minimax distance

$$\mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) = \sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}}\delta(f, g),$$

with respect to some metric δ.

Proposition 3.1. *Let $(\mathcal{X}, \delta)$ be a metric space, $\mathcal{F}, \mathcal{G} \subseteq \mathcal{X}$, and $\varepsilon \in \mathbb{R}_+$. Suppose that*

$$\mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) \leq \varepsilon. \tag{39}$$

Then,

$$N(\varepsilon, \mathcal{F}, \delta) \geq N(4\varepsilon, \mathcal{G}, \delta). \tag{40}$$

Proof. See Appendix B.1. □

3.1 Neural Network Transformation

Generally speaking, neural network transformation is the practice of approximating or exactly realizing a given neural network with certain structural properties by another neural network satisfying different prescribed structural properties. This problem has a number of concrete incarnations. For example, in network compression the objective is to reduce the size of networks. In practice, this is often effected through techniques such as pruning [26, 27] or knowledge distillation [28]. Another example is network quantization, where real-valued network weights are replaced by weights that are quantized to a predetermined level of precision, or high-precision weights are substituted by lower-precision weights. This can be done either

by rounding each individual weight to the nearest quantization point or by searching for the best set of quantized weights jointly through specific algorithms [29]. The primary motivation for network compression and quantization stems from the necessity to store neural networks on microchips under prescribed memory constraints. Further examples of neural network transformation appear in [30, Theorem 3.1] where a given network is transformed into one that is narrower and deeper, and in [10, Lemma A.1], [13, Theorem 5], [30, Corollary 3.2] which all employ transformations into networks of smaller weight magnitude.

More formally, the problem of neural network transformation can be cast as follows. Considering the classes of networks $\mathcal{R}_1$ and $\mathcal{R}_2$, one wants to approximate a given network $r_1 \in \mathcal{R}_1$ by a network $r_2 \in \mathcal{R}_2$ such that the distance $\delta(r_1, r_2)$, for some metric δ, is minimized. The fundamental limit on the worst-case error incurred by the transformation mapping $\mathfrak{C} : \mathcal{R}_1 \mapsto \mathcal{R}_2$, under the metric δ, is characterized by the minimax approximation error $\mathcal{A}(\mathcal{R}_1, \mathcal{R}_2, \delta)$ according to

$$\sup_{r_1 \in \mathcal{R}_1} \delta(r_1, \mathfrak{C}(r_1)) \geq \sup_{r_1 \in \mathcal{R}_1} \inf_{r_2 \in \mathcal{R}_2} \delta(r_1, r_2) = \mathcal{A}(\mathcal{R}_1, \mathcal{R}_2, \delta).$$

In [31, Theorem 1.1], for example, a lower bound on $\mathcal{A}(\mathcal{R}_1, \mathcal{R}_2, \delta)$ was provided in terms of the oscillation count of ReLU networks (as defined in [31, Sec. 3]), in the case where deep networks are replaced by shallow ones. We next show how $\mathcal{A}(\mathcal{R}_1, \mathcal{R}_2, \delta)$ can be characterized for general $\mathcal{R}_1$ and $\mathcal{R}_2$ through covering numbers. For concreteness, we consider $\mathcal{R}_1 = \mathcal{R}(d, W, L, B), \mathcal{R}_2 = \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B})$ with $\delta = L^p([0,1]^d)$ for general $p \in [1, \infty]$.

Corollary 3.2. *Let $p \in [1, \infty]$, $d \in \mathbb{N}$, $L, \widetilde{L}, W, \widetilde{W} \in \mathbb{N}$, with $W, L \geq 60$, $B, \widetilde{B} \in \mathbb{R}_+$, with $B, \widetilde{B} \geq 1$. Assume that there exists an $\varepsilon \in (0, 1/8)$ such that*

$$\mathcal{A}(\mathcal{R}(d, W, L, B), \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B}), \| \cdot \|_{L^p([0,1]^d)}) \leq \varepsilon. \tag{41}$$

Then,

$$c\, W^2 L \log\left(\frac{(W+1)^L B^L}{4\varepsilon} \right) \leq C\, \widetilde{W}^2 \widetilde{L} \log\left(\frac{(\widetilde{W}+1)^{\widetilde{L}} \widetilde{B}^{\widetilde{L}}}{\varepsilon} \right), \tag{42}$$

where C and c are the absolute constants in Theorem 2.1 corresponding to the parameters $\widetilde{W}, \widetilde{L}, \widetilde{B}$ and W, L, B, respectively. In particular, if $\mathcal{R}(d, W, L, B) \subseteq \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B})$, we have

$$C\, \widetilde{W}^2 \widetilde{L} \geq c\, W^2 L. \tag{43}$$

Proof. Application of Proposition 3.1 with $\delta = \| \cdot \|_{L^p([0,1]^d)}$, $\mathcal{G} = \mathcal{R}(d, W, L, B)$, and $\mathcal{F} = \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B})$, and the prerequisite (39) satisfied thanks to (41), yields

$$N(\varepsilon, \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B}), L^p([0,1]^d)) \geq N(4\varepsilon, \mathcal{R}(d, W, L, B), L^p([0,1]^d)),$$

which together with Theorem 2.1 establishes (42). If $\mathcal{R}(d, W, L, B) \subseteq \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B})$, then (41) holds for all $\varepsilon \in \left(0, \frac{1}{8}\right)$ and consequently so does (42). Dividing (42) by $\log(\frac{1}{\varepsilon})$ and letting $\varepsilon \to 0$ results in (43). $\qquad\square$

Corollary 3.2 allows us to answer the following question on network size reduction: Is it possible to approximate a network in $\mathcal{N}(d, W, L, B)$ by one in $\mathcal{N}(d, \widetilde{W}, \widetilde{L}, \widetilde{B})$, with prescribed error ε, while having the maximum number of nonzero weights of the approximating network, $\widetilde{W}^2 \widetilde{L}$, be of order smaller than that of the original network? When $\varepsilon = 0$, i.e., $\mathcal{R}(d, W, L, B) \subseteq \mathcal{R}(d, \widetilde{W}, \widetilde{L}, \widetilde{B})$, (43) shows that the answer is negative. For $\varepsilon \in (0, 1/8)$, we can conclude from (42) that this would require that the weight magnitude $\widetilde{B}$ compensate for the reduction in an exponential manner.

We next consider network transformation through weight quantization.

Corollary 3.3. *Let $p \in [1, \infty]$, $d, W, L \in \mathbb{N}$, $B \in \mathbb{R}_+$, with $B \geq 1$ and $W, L \geq 60$. Let $\mathbb{A} \subseteq \mathbb{R}$ be a finite set such that $|\mathbb{A}| \geq 2$. Then, we have*

$$\mathcal{A}(\mathcal{R}(d, W, L, B), \mathcal{R}_{\mathbb{A}}(d, W, L), \|\cdot\|_{L^p([0,1]^d)}) \geq \min\{1/8, (W+1)^L B^L 2^{-c \log(|\mathbb{A}|)}\}, \qquad (44)$$

for some absolute constant $c \in \mathbb{R}_+$.

Proof. Let
$$\kappa := \mathcal{A}(\mathcal{R}(d, W, L, B), \mathcal{R}_{\mathbb{A}}(d, W, L), \|\cdot\|_{L^p([0,1]^d)}). \qquad (45)$$

When $\kappa \geq 1/8$, the desired inequality (44) holds trivially. For $\kappa < 1/8$, it follows from Proposition 3.1 with $\varepsilon = \kappa$, $\delta = \|\cdot\|_{L^p([0,1]^d)}$, $\mathcal{G} = \mathcal{R}(d, W, L, B)$, and $\mathcal{F} = \mathcal{R}_{\mathbb{A}}(d, W, L)$ that

$$N(\kappa, \mathcal{R}_{\mathbb{A}}(d, W, L), L^p([0,1]^d)) \geq N(4\kappa, \mathcal{R}(d, W, L, B), L^p([0,1]^d)). \qquad (46)$$

We next note that

$$\log(N(\kappa, \mathcal{R}_{\mathbb{A}}(d, W, L), L^p([0,1]^d))) \leq \log(|\mathcal{R}_{\mathbb{A}}(d, W, L)|) \qquad (47)$$
$$\leq 5W^2 L \log(|\mathbb{A}|), \qquad (48)$$

where (47) follows from the fact that every set is a covering of itself and (48) is by Lemma 2.4. Further, it follows from (4) in Theorem 2.1 with $\varepsilon = 4\kappa$, that

$$\log(N(4\kappa, \mathcal{R}(d, W, L, B), L^p([0,1]^d))) \geq c_1 W^2 L \log\left(\frac{(W+1)^L B^L}{4\kappa}\right), \qquad (49)$$

with $c_1 \in \mathbb{R}_+$ an absolute constant. Using (47)-(48) and (49) in (46), yields

$$5W^2 L \log(|\mathbb{A}|) \geq c_1 W^2 L \log\left(\frac{(W+1)^L B^L}{4\kappa}\right),$$

which implies $\kappa \geq \frac{1}{4}(W+1)^L B^L 2^{-\frac{5}{c_1} \log(|\mathbb{A}|)} \geq (W+1)^L B^L 2^{-\left(\frac{5}{c_1}+2\right) \log(|\mathbb{A}|)}$, thanks to $\log(|\mathbb{A}|) \geq 1$. The proof is concluded by setting $c := \frac{5}{c_1} + 2$. $\qquad \square$

Corollary 3.3 allows us to conclude that the worst-case quantization error $\mathcal{A}(\mathcal{R}(d, W, L, B),$ $\mathcal{R}_{\mathbb{A}}(d, W, L), \|\cdot\|_{L^p([0,1]^d)})$ decreases no faster than exponential in the number of bits $\log(|\mathbb{A}|)$ required to store the elements of $\mathbb{A}$. Moreover, as W, L, and B grow, the network weight resolution has to increase in order to compensate for the growth in the factor $(W + 1)^L B^L$. Specifically, if we require that $\mathcal{A}(\mathcal{R}(d, W, L, B), \mathcal{R}_{\mathbb{A}}(d, W, L), \|\cdot\|_{L^p([0,1]^d)}) \leq \kappa$, we must have $\log(|\mathbb{A}|) \geq \frac{1}{c} \log\left(\frac{(W+1)^L B^L}{\kappa}\right)$. This lower bound can be achieved, within a multiplicative constant, by taking $\mathbb{A}$ to be an equidistant set contained in the interval $[-B, B]$. To see this, we set $\mathbb{A} = [-B, B] \cap 2^{-b}\mathbb{Z}$ with $b = \lceil \log(\frac{L(W+1)^L B^{L-1}}{\kappa}) \rceil$ and note that

$$\mathcal{A}(\mathcal{R}(d, W, L, B), \mathcal{R}_{[-B,B] \cap 2^{-b}\mathbb{Z}}(d, W, L), \|\cdot\|_{L^p([0,1]^d)}) \qquad (50)$$
$$\leq L(W+1)^L B^{L-1} 2^{-b} \qquad (51)$$
$$\leq \kappa, \qquad (52)$$

where in (51) we applied Lemma 2.3. The argument is concluded upon realizing that $\log(|\mathbb{A}|) = \log(|[-B, B] \cap 2^{-b}\mathbb{Z}|) \leq 3 \log(\frac{L(W+1)^L B^L}{\kappa}) \leq 6 \log(\frac{(W+1)^L B^L}{\kappa})$, where the first inequality follows from (13)-(18) and the second is by $\frac{L(W+1)^L B^L}{\kappa} \leq \frac{(W+1)^L \cdot (W+1)^L B^L}{\kappa} \leq (\frac{(W+1)^L B^L}{\kappa})^2$.

We finally emphasize that the results on the fundamental limits on neural network transformation presented in this section are made possible by the tight covering number lower bound (4) in Theorem 2.1.

3.2 Deep Neural Network Function Approximation

We next show how the covering number bounds in Theorem 2.1, in combination with Proposition 3.1, can be used to establish a tight characterization of the minimax error in the ReLU network approximation of the class of 1-Lipschitz functions

$$H^1([0,1]) := \{f \in C([0,1]) : |f(x)| \leq 1, |f(x) - f(y)| \leq |x - y|, \forall x, y \in [0,1]\}. \quad (53)$$

To this end, we start with the following upper bound on the minimax error.

Lemma 3.4. *There exist absolute constants $C, D \in \mathbb{R}_+$ such that, for all $W, L \in \mathbb{N}$, with $W, L \geq D$, and $p \in [1, \infty]$,*

$$\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^p([0,1])}) \leq C(W^2 L^2 \log(W))^{-1}. \quad (54)$$

Proof. By [18, Theorem 3.1], there exist absolute constants $C, D \in \mathbb{R}_+$ such that, for all $W, L \in \mathbb{N}$, with $W, L \geq D$, $\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^\infty([0,1])}) \leq C(W^2 L^2 \log(W))^{-1}$. Noting that, for all $p \in [1, \infty]$, the $L^p([0,1])$-norm is dominated by the $L^\infty([0,1])$-norm, i.e., $\|f\|_{L^p([0,1])} \leq \|f\|_{L^\infty([0,1])}, \forall f \in L^\infty([0,1])$, we have, for all $W, L \in \mathbb{N}$, with $W, L \geq D$, and $p \in [1, \infty]$, $\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^p([0,1])}) \leq \mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^\infty([0,1])}) \leq C(W^2 L^2 \log(W))^{-1}$. $\square$

A corresponding lower bound, for $p = \infty$ and $L \geq 2$, obtained through arguments involving VC dimension is given by [18, Proposition 2.11],

$$\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^\infty([0,1])}) \geq c_v(W^2 L^2 (\log(W) + \log(L)))^{-1}, \quad (55)$$

where c_v is an absolute constant. A comparison of the upper bound (54) and the lower bound (55) reveals a gap owing to the additive term $\log(L)$ in the lower bound. The question is now whether the lower or the upper bound would need to be refined to close this gap. The tight covering number bounds in Theorem 2.1 allow to answer this question. Concretely, it turns out that it is the lower bound that can be strengthened. The resulting improvement pertains to all $p \in [1, \infty]$. To see all this, we start with a lower bound on the covering number of $H^1([0,1])$.

Lemma 3.5. *There exists an absolute constant $C \in \mathbb{R}_+$ such that, for all $p \in [1, \infty]$ and $\varepsilon \in (0, 1/2)$,*

$$\log(N(\varepsilon, H^1([0,1]), L^p([0,1]))) \geq C\varepsilon^{-1}. \quad (56)$$

For $p = \infty$, the statement of Lemma 3.5 is [6, Example 5.10]. An asymptotic version of Lemma 3.5 was provided in [32, Theorem 1.7]. Inspection of the proof of [32, Theorem 1.7] reveals quite directly that the result holds in nonasymptotic form as stated here and does so even for a more general class of functions. The proof of Lemma 3.5 is hence omitted.

We are now ready to present the strengthened lower bound.

Corollary 3.6. *Let $p \in [1, \infty]$, $W, L \in \mathbb{N}$. It holds that*

$$\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^p([0,1])}) \geq \min\left\{\frac{1}{8}, c\,(W^2 L^2 \log(W))^{-1}\right\}, \quad (57)$$

where $c \in \mathbb{R}_+$ is an absolute constant.

Proof. Let

$$\kappa := \mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^p([0,1])}). \quad (58)$$

When $\kappa \geq 1/8$, the desired inequality (57) holds trivially. For $\kappa < 1/8$, it follows from Proposition 3.1 with $\varepsilon = \kappa$, $\delta = \|\cdot\|_{L^p([0,1])}$, $\mathcal{F} = \mathcal{R}(1, W, L, 1)$, and $\mathcal{G} = H^1([0,1])$ that

$$N(\kappa, \mathcal{R}(1, W, L, 1), L^p([0,1])) \geq N(4\kappa, H^1([0,1]), L^p([0,1])). \tag{59}$$

The left-hand-side of (59) can now be upper-bounded by (3) in Theorem 2.1 according to

$$\log(N(\kappa, \mathcal{R}(1, W, L, 1), L^p([0,1]))) \leq C_1 W^2 L \log\left(\frac{(W+1)^L}{\kappa}\right), \tag{60}$$

with C_1 an absolute constant. Application of Lemma 3.5 with $\varepsilon = 4\kappa$, yields

$$\log(N(4\kappa, H^1([0,1]), L^p([0,1]))) \geq C_2 \kappa^{-1}, \tag{61}$$

with C_2 an absolute constant. Using (60) and (61) in (59) leads to

$$C_1 W^2 L^2 \log(W+1) + C_1 W^2 L \log(\kappa^{-1}) - C_2 \kappa^{-1} \geq 0. \tag{62}$$

Next, we define $f : \mathbb{R} \mapsto \mathbb{R}$ as

$$f(x) = C_1 W^2 L^2 \log(W+1) + C_1 W^2 L \log(x) - C_2 x, \tag{63}$$

which allows us to rewrite (62) as

$$f(\kappa^{-1}) \geq 0. \tag{64}$$

We proceed to characterize the feasible set $\{x : f(x) \geq 0\}$ for κ^{-1}. First, note that the set $\{x \in [100, \infty) : 2^x - \frac{2C_1}{C_2} \cdot (x + 10) \geq 0\}$ is nonempty, and let $\nu := \inf\{x \in [100, \infty) : 2^x - \frac{2C_1}{C_2} \cdot (x + 10) \geq 0\} \in [100, \infty)$. Thanks to the continuity of the mapping $x \in \mathbb{R} \mapsto 2^x - \frac{2C_1}{C_2} \cdot (x + 10)$, we have $2^\nu - \frac{2C_1}{C_2} \cdot (\nu + 10) \geq 0$. Moreover, ν depends on $\frac{C_1}{C_2}$ only. Let

$$b := \frac{2C_1}{C_2} \cdot (\nu + 10).$$

Then, we have $2^\nu \geq b$, and

$$f\left(b W^2 L^2 \log(W+1)\right) \tag{65}$$

$$= C_1 W^2 L^2 \log(W+1) + C_1 W^2 L \log(b W^2 L^2 \log(W+1)) - C_2 \, b \, W^2 L^2 \log(W+1) \tag{66}$$

$$= \left(C_1 - \frac{C_2 b}{2}\right) W^2 L^2 \log(W+1)$$

$$\quad + C_1 W^2 L \left(\log(b W^2 L^2 \log(W+1)) - \frac{C_2 b}{2 C_1} L \log(W+1)\right) \tag{67}$$

$$< C_1 W^2 L \left(\log\left(b W^2 L^2 \log(W+1)\right) - \log\left((W+1)^{\frac{C_2 b}{2C_1} L}\right)\right) \tag{68}$$

$$\leq 0, \tag{69}$$

where (68) follows from $C_1 - \frac{C_2 b}{2} = C_1 - \frac{C_2}{2} \cdot \frac{2C_1}{C_2} \cdot (\nu + 10) < C_1 - \frac{C_2}{2} \cdot \frac{2C_1}{C_2} = 0$, and in (69) we used

$$(W+1)^{\frac{C_2 b}{2C_1} L} = (W+1)^{(\frac{C_2 b}{2C_1} - 10)L} (W+1)^{10L} \tag{70}$$

$$\geq 2^{\frac{C_2 b}{2C_1} - 10} (W+1)^2 \cdot ((W+1)^L)^2 \cdot (W+1) \tag{71}$$

$$\geq 2^{\frac{C_2 b}{2C_1} - 10} W^2 L^2 \log(W+1) \tag{72}$$

$$= 2^\nu W^2 L^2 \log(W+1) \tag{73}$$

$$\geq b W^2 L^2 \log(W+1). \tag{74}$$

13

We next note that f is strictly decreasing on $[bW^2L^2\log(W+1),\infty)$ as $f'(x) = \frac{C_1W^2L}{x\ln(2)} - C_2 \leq$ $\frac{C_1W^2L}{bW^2L^2\log(W+1)\ln(2)} - C_2 \leq \frac{C_1}{b\ln(2)} - C_2 = \frac{C_1}{\frac{2C_1}{C_2}\cdot(\nu+10)\ln(2)} - C_2 < 0$, for all $x \in [bW^2L^2\log(W+1),\infty)$. It hence follows from (65)-(69) that $f(x) < 0$, for all $x > bW^2L^2\log(W+1)$, and therefore

$$\{x : f(x) \geq 0\} \subseteq (-\infty, bW^2L^2\log(W+1)]. \tag{75}$$

Putting (64) and (75) together, we obtain

$$\kappa^{-1} \leq bW^2L^2\log(W+1), \tag{76}$$

which, in turn, implies

$$\kappa \geq b^{-1}(W^2L^2\log(W+1))^{-1} \geq \min\left\{\frac{1}{8}, b^{-1}(W^2L^2\log(W+1))^{-1}\right\}.$$

The proof is concluded by setting $c = b^{-1}$. $\qquad\square$

4 Optimal Rates in Nonparametric Regression

In this section, we show how the minimax error upper bound in Lemma 3.4 leads to a sharp characterization of the prediction error in nonparametric regression through ReLU networks. The general results we obtain allow to infer, inter alia, that nonparametric regression with very deep[2] fully-connected ReLU networks achieves optimal sample complexity rate in the estimation of 1-Lipschitz functions; this improves significantly upon [8, Theorem 1(b)] in the special case of 1-Lipschitz functions by removing the $(\log(n))^6$-factor. The section concludes with insights on a systematic relation between optimal nonparametric regression and optimal approximation through (deep) ReLU networks unifying numerous corresponding results in the literature [13, 14, 21] and identifying general underlying principles.

The goal of nonparametric regression is to estimate the unknown function $g : \mathbb{X} \mapsto \mathbb{R}$, with $\mathbb{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, referred to as the regression function, from the $n \in \mathbb{N}$ (random) samples

$$(x_i, y_i)_{i=1}^n = (x_i, g(x_i) + \sigma\xi_i)_{i=1}^n, \tag{77}$$

where $\sigma \in \mathbb{R}_+$, $(x_i)_{i=1}^n$ are i.i.d. random variables of distribution P supported on $\mathbb{X}$, $(\xi_i)_{i=1}^n$ are i.i.d. standard (i.e., zero mean and unit variance) Gaussian random variables, and $(x_i)_{i=1}^n$ and $(\xi_i)_{i=1}^n$ are statistically independent.

4.1 Nonparametric Regression through ReLU Networks

Nonparametric regression through ReLU networks was considered in [8, 13, 14, 21], with g estimated by fitting a network $\hat{f}_n$ from a given class $\mathcal{F}_n$ of networks through minimization of the empirical risk $\frac{1}{n}\sum_{i=1}^n(\hat{f}_n(x_i) - y_i)^2$. For example, [13] considers regression functions g that can be written as the composition of bounded Hölder functions and $\mathcal{F}_n$ is a family of sparse ReLU networks with bounded output. The quality of the estimator is generally measured by the so-called prediction error

$$\|\hat{f}_n - g\|^2_{L^2(P)} = \int |\hat{f}_n(x) - g(x)|^2\, dP(x).$$

---

[2]Here, "very deep" refers to networks whose depth increases at least linearly in network width, which is in contrast to networks commonly considered in the literature [10, 13, 14, 21] that have depth increasing at most logarithmically in width.

The references [8, 13, 14, 21] report upper bounds on the prediction error. Notably, the bounds in [8] are derived employing arguments based on VC-dimension, while those in [13, 14, 21] are obtained from covering number upper bounds for ReLU networks. The following Theorem 4.1 summarizes the results [13, Lemma 4], [14, Lemma 4 and Lemma 5] and reformulates them so as to highlight the individual effects of the approximation error and the covering number of $\mathcal{F}_n$. Application of the minimax error upper bound in Lemma 3.4 then results in the removal of the $(\log(n))^6$-factor in the special case of Lipschitz functions in [8, Theorem 1(b)]. Moreover, our reformulation sets the stage for the development of a fundamental relation between optimal approximation and optimal regression through ReLU networks provided at the end of this section. We emphasize that most of the techniques and ideas used in the proof of Theorem 4.1 follow [13].

Theorem 4.1. *Let $\mathbb{X} \subseteq \mathbb{R}^d$ and consider the regression function $g : \mathbb{X} \mapsto \mathbb{R}$. Let $n \in \mathbb{N}$ and $\sigma \in \mathbb{R}_+$. Let P be a distribution on $\mathbb{X}$, with the associated samples $(x_i, y_i)_{i=1}^n = (x_i, g(x_i) + \sigma\xi_i)_{i=1}^n$, where $(x_i)_{i=1}^n$ are i.i.d. random variables of distribution P, $(\xi_i)_{i=1}^n$ are i.i.d. standard Gaussian random variables, and $(x_i)_{i=1}^n$ and $(\xi_i)_{i=1}^n$ are statistically independent.*

Let $\varepsilon \in (0, 1/2)$, and consider a class of functions $\mathcal{F}_n \subseteq L^\infty(\mathbb{X})$ such that

$$\inf_{f \in \mathcal{F}_n} \|g - f\|_{L^2(P)} \leq \varepsilon \tag{78}$$

and an $\mathcal{F}_n$-valued random variable $\hat{f}_n$ satisfying

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - y_i)^2 \leq \inf_{f \in \mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2\right) + \varepsilon^2, \quad a.s. \tag{79}$$

It holds that

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \leq C(1 + \sigma + \sigma^2 + (R(g, \mathcal{F}_n))^2)\left(\varepsilon^2 + \frac{\log(N(\varepsilon^2, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}\right), \tag{80}$$

where $C \in \mathbb{R}_+$ is an absolute constant and $R(g, \mathcal{F}_n) := \max\{\|g\|_{L^\infty(\mathbb{X})}, \sup_{f \in \mathcal{F}_n} \|f\|_{L^\infty(\mathbb{X})}\}$.

Proof. See Appendix C. □

The prediction error upper bound in Theorem 4.1 relies on two assumptions. The first one is the approximation assumption (78), which states that the regression function g can be approximated well by functions in $\mathcal{F}_n$. The second is the empirical risk minimization assumption (79), which requires $\hat{f}_n$, almost surely, to nearly achieve the minimal empirical risk among $\mathcal{F}_n$.

We proceed to apply Theorem 4.1 to the estimation of 1-Lipschitz functions $g \in H^1([0, 1])$ from the associated (random) samples $(x_i, g(x_i) + \sigma\xi_i)_{i=1}^n$ using nonparametric least squares with very deep fully-connected ReLU networks of fixed width and output truncated to the interval $[-1, 1]$. Formally, the truncation is effected by applying the operator $\mathcal{T}_E : \mathbb{R} \mapsto [-E, E]$, $E \in \mathbb{R}_+$,

$$\mathcal{T}_E(x) = \max\{-E, \min\{E, x\}\}, \tag{81}$$

with $E = 1$ to the neural network output. This truncation is commonly adopted in the literature [8, 13, 14, 21] and it is quite natural given that the regression function g to be estimated satisfies $\|g\|_{L^\infty([0,1])} \leq 1$. The formal result can now be stated as follows.

Corollary 4.2. *Consider the regression function $g \in H^1([0, 1])$. Let $n \in \mathbb{N}$ and $\sigma \in \mathbb{R}_+$. Let P be a distribution on $[0, 1]$, with the associated samples $(x_i, y_i)_{i=1}^n = (x_i, g(x_i) + \sigma\xi_i)_{i=1}^n$, where $(x_i)_{i=1}^n$ are i.i.d. random variables of distribution P, $(\xi_i)_{i=1}^n$ are i.i.d. standard Gaussian random variables, and $(x_i)_{i=1}^n$ and $(\xi_i)_{i=1}^n$ are statistically independent.*

Let C and D be the constants specified in Lemma 3.4, and set

$$L(n) := \lceil 2(D+1)(C+1)^{1/2}n^{1/6}\rceil, \quad \mathcal{F}_n := \mathcal{T}_1 \circ \mathcal{R}(1, \lceil D+1\rceil, L(n), 1). \tag{82}$$

Let $\hat{f}_n$ be the empirical risk minimizer[3] in $\mathcal{F}_n$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(x_i) - y_i)^2 = \inf_{f\in\mathcal{F}_n}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2\right), \quad a.s. \tag{83}$$

Then,

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \le K(\sigma)n^{-2/3}, \tag{84}$$

where $K(\sigma)$ is a constant depending on σ only.

Proof. We apply Theorem 4.1 with $\mathbb{X} = [0,1]$ and choose $\varepsilon \in (0, 1/2)$ such that the prerequisites (78) and (79) are satisfied. To this end, we first apply Lemma 3.4 with $W = \lceil D+1\rceil \ge D$, $L = L(n) = \lceil 2(D+1)(C+1)^{1/2}n^{1/6}\rceil \ge D$, and $p = \infty$ to obtain

$$
\begin{aligned}
\mathcal{A}(H^1([0,1]), \mathcal{R}(1, \lceil D+1\rceil, L(n), 1), \|\cdot\|_{L^\infty([0,1])}) &\le C(\lceil D+1\rceil^2(L(n))^2\log(\lceil D+1\rceil))^{-1} \\
&\le C(L(n))^{-2} \\
&= C(\lceil 2(D+1)(C+1)^{1/2}n^{1/6}\rceil)^{-2} \\
&\le \frac{1}{4}n^{-1/3}.
\end{aligned} \tag{85}
$$

It then follows that

$$
\begin{aligned}
\inf_{f\in\mathcal{F}_n}\|f - g\|_{L^2(P)} &= \inf_{f\in\mathcal{R}(1,\lceil D+1\rceil,L(n),1)}\|\mathcal{T}_1\circ f - g\|_{L^2(P)} \tag{86} \\
&\le \inf_{f\in\mathcal{R}(1,\lceil D+1\rceil,L(n),1)}\|\mathcal{T}_1\circ f - g\|_{L^\infty([0,1])} \tag{87} \\
&= \inf_{f\in\mathcal{R}(1,\lceil D+1\rceil,L(n),1)}\|\mathcal{T}_1\circ f - \mathcal{T}_1\circ g\|_{L^\infty([0,1])} \tag{88} \\
&\le \inf_{f\in\mathcal{R}(1,\lceil D+1\rceil,L(n),1)}\|f - g\|_{L^\infty([0,1])} \tag{89} \\
&\le \mathcal{A}(H^1([0,1]), \mathcal{R}(1, \lceil D+1\rceil, L(n), 1), \|\cdot\|_{L^\infty([0,1])}) \tag{90} \\
&\le \frac{1}{4}n^{-1/3}, \tag{91}
\end{aligned}
$$

where in (87) we used that P is a distribution on $[0,1]$, (88) follows from the fact that $g \in H^1([0,1])$ takes values in $[-1,1]$, (89) is a consequence of $\mathcal{T}_1$ being 1-Lipschitz, and in (91) we employed (85). We have therefore verified (78) with

$$\varepsilon := \frac{1}{4}n^{-1/3}.$$

Prerequisite (79) holds with the same $\varepsilon = \frac{1}{4}n^{-1/3}$ owing to assumption (83). We are now in a position to apply Theorem 4.1 resulting in

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \le C_1(1 + \sigma + \sigma^2 + (R(g, \mathcal{F}_n))^2)\left(\varepsilon^2 + \frac{\log(N(\varepsilon^2, \mathcal{F}_n, L^\infty([0,1]))) + 1}{n}\right) \tag{92}$$

$$\le C_1(2 + \sigma + \sigma^2)\left(\frac{1}{16}n^{-2/3} + \frac{\log(N(\frac{1}{16}n^{-2/3}, \mathcal{F}_n, L^\infty([0,1]))) + 1}{n}\right), \tag{93}$$

---

[3]The existence of the minimizer is argued in Section F.3. For simplicity of exposition, we assume that the minimizer can be identified exactly, thereby ignoring the impact of suboptimality of the optimization algorithm employed. This simplification is common in the literature, see e.g. [8, 21]. We note, however, that Theorem 4.1 can accommodate cases where minimization is accomplished only approximately.

where C_1 is the absolute constant C from Theorem 4.1, and in (93) we used $R(g, \mathcal{F}_n) \leq 1$ which follows from $g \in H^1([0,1])$ and the fact that $\mathcal{F}_n$ consists of functions that take values in $[-1, 1]$.

We next upper-bound the term $\log(N(\frac{1}{16}n^{-2/3}, \mathcal{F}_n, L^\infty([0,1])))$. As $\mathcal{F}_n = \mathcal{T}_1 \circ \mathcal{R}(1, \lceil D + 1 \rceil, L(n), 1)$ and $\mathcal{T}_1$ is 1-Lipschitz, every ε-covering $\{x_i\}_{i=1}^N$ of $\mathcal{R}(1, \lceil D+1 \rceil, L(n), 1)$ with respect to the $L^\infty([0,1])$-norm induces an ε-covering $\{\mathcal{T}_1 \circ x_i\}_{i=1}^N$ of $\mathcal{F}_n$ with respect to the $L^\infty([0,1])$-norm. It therefore holds that

$$N\left(\frac{1}{16}n^{-2/3}, \mathcal{F}_n, L^\infty([0,1])\right) \leq N\left(\frac{1}{16}n^{-2/3}, \mathcal{R}(1, \lceil D + 1 \rceil, L(n), 1), L^\infty([0,1])\right). \tag{94}$$

The right-hand-side of (94) can now be upper-bounded according to

$$\log\left(N\left(\frac{1}{16}n^{-2/3}, \mathcal{R}(1, \lceil D + 1 \rceil, L(n), 1), L^\infty([0,1])\right)\right) \tag{95}$$

$$\leq C_2\lceil D + 1 \rceil^2 L(n) \log\left(\frac{(\lceil D + 1 \rceil + 1)^{L(n)}}{\frac{1}{16}n^{-2/3}}\right) \tag{96}$$

$$= C_2\lceil D + 1 \rceil^2 L(n)\left(L(n)\log(\lceil D + 1 \rceil + 1) + \log(16\,n^{2/3})\right) \tag{97}$$

$$\leq C_2\lceil D + 1 \rceil^2 L(n)\left(L(n)\log(\lceil D + 1 \rceil + 1) + C_3 n^{1/6}\right) \tag{98}$$

$$\leq C_2\lceil D + 1 \rceil^2 C_4 n^{1/6}(C_4 n^{1/6} \log(\lceil D + 1 \rceil + 1) + C_3 n^{1/6}) \tag{99}$$

$$= C_5 n^{1/3}, \tag{100}$$

where (96) follows by application of (3) in Theorem 2.1 with $p = \infty$, $\varepsilon = \frac{1}{16}n^{-2/3}$, $d = 1$, $W = \lceil D + 1 \rceil$, $L = L(n)$, and $B = 1$, and C_2 is the absolute constant C from Theorem 2.1, in (98) we set $C_3 := \sup_{x \in [1,\infty)} \frac{\log(16 x^{2/3})}{x^{1/6}} < \infty$, which is an absolute constant, in (99) we used $L(n) = \lceil 2(D+1)(C+1)^{1/2} n^{1/6} \rceil \leq 2(D+1)(C+1)^{1/2} n^{1/6} + 1 \leq 4(D+1)(C+1)^{1/2} n^{1/6}$ and let $C_4 := 4(D+1)(C+1)^{1/2}$, and in (100) we set $C_5 = C_2\lceil D + 1 \rceil^2 C_4(C_4 \log(\lceil D + 1 \rceil + 1) + C_3)$. Using (94) and (95)-(100) in (92)-(93), finally yields

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \leq C_1(2 + \sigma + \sigma^2)\left(\frac{1}{16}n^{-2/3} + \frac{C_5 n^{1/3} + 1}{n}\right)$$

$$\leq C_1(2 + \sigma + \sigma^2)\left(\frac{1}{16}n^{-2/3} + \frac{(C_5 + 1)n^{1/3}}{n}\right)$$

$$= C_1(2 + \sigma + \sigma^2)\left(\frac{1}{16} + C_5 + 1\right)n^{-2/3}.$$

The proof is finalized by taking $K(\sigma) = C_1(2 + \sigma + \sigma^2)(\frac{1}{16} + C_5 + 1)$. $\qquad\square$

The rate $n^{-2/3}$ in Corollary 4.2 is optimal [33, Theorem 1], see also [34, Theorem 3.2] and [13, Theorem 3]. In particular, compared to the corresponding best known result in the literature given by an upper bound of rate $(\log(n))^6 n^{-2/3}$ [8, Theorem 1(b)], Corollary 4.2 disposes of the $(\log(n))^6$-factor. We note that [8, Theorem 1(b)] applies to a more general class of smooth functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. The removal of the $(\log(n))^6$-factor carries through to Lipschitz functions on $\mathbb{R}^d$ with general $d \in \mathbb{N}$, but we do not present the details here. The improvement we obtain stems from the approximation result Lemma 3.4 and the use of the covering number instead of VC-dimension as in [8]. More specifically, in the approximation of functions in $H^1([0,1])$ through very deep fully-connected ReLU networks of fixed width and depth L, both our Lemma 3.4 and [8, Theorem 2(b)] achieve guaranteed error decay of L^{-2}. However, [8, Theorem 2(b)] requires networks with arbitrarily large weight-magnitude, corresponding to

unbounded sets, whereas Lemma 3.4 needs networks of weight magnitude bounded by 1 only. This significant reduction in the size of the model class, to compact sets, makes it possible to upper-bound the prediction error through covering numbers.

4.2 Optimal Regression and Optimal Approximation

The optimality established in Corollary 4.2 in the previous section is a consequence of a deeper set of ideas, which we now bring to the fore and develop in a more general context. This discussion will demonstrate how optimality in function approximation through ReLU networks along with their covering number behavior plays a fundamental role in attaining optimal regression. As a byproduct, we will be able to shed light on the specific choices for $L(n)$ and $\mathcal{F}_n$ in Corollary 4.2.

We build on the information-theoretic characterization of optimal sample complexity rates developed by Yang and Barron [35]. Concretely, it is shown in [35, Section 3.2] that, for a uniformly bounded function class $\mathcal{G}$, the optimal sample complexity rate in the estimation of regression functions $g \in \mathcal{G}$ is determined by the covering number of $\mathcal{G}$. To make these results more concrete, consider the general nonparametric regression setup introduced at the beginning of Section 4. The estimation of $g \in \mathcal{G}$ from random samples can now be described as the application of a mapping $\mathfrak{F}_n : (\mathbb{X} \times \mathbb{R})^n \mapsto L^\infty(\mathbb{X})$ that takes the samples $(x_i, y_i)_{i=1}^n$ to the estimate $\hat{f}_n \in \mathcal{G}$. For example, in Corollary 4.2, $\mathfrak{F}_n((x_i, y_i)_{i=1}^n)$ would be the mapping induced by the empirical risk minimizer defined according to (83). By [35, Theorem 6], under a weak technical condition[4] on the packing number of $\mathcal{G}$, it holds that, for all $\mathfrak{F}_n : (\mathbb{X} \times \mathbb{R})^n \mapsto L^\infty(\mathbb{X})$,

$$\sup_{g \in \mathcal{G}} E(\|\mathfrak{F}_n((x_i, g(x_i) + \sigma\xi_i)_{i=1}^n) - g\|_{L^2(P)}^2) \geq c(\mathcal{G}, \sigma, P)\kappa_n^2, \tag{101}$$

where κ_n is the solution to the equation

$$\kappa_n^2 = \frac{\log(M(\kappa_n, \mathcal{G}, L^2(P)))}{n}, \tag{102}$$

and $c(\mathcal{G}, \sigma, P) \in \mathbb{R}_+$ is a constant depending on $\mathcal{G}, \sigma$, and P only. The lower bound (101) is achievable in the sense of the existence of an $\mathfrak{F}_n : (\mathbb{X} \times \mathbb{R})^n \mapsto L^\infty(\mathbb{X})$ such that

$$\sup_{g \in \mathcal{G}} E(\|\mathfrak{F}_n((x_i, g(x_i) + \sigma\xi_i)_{i=1}^n) - g\|_{L^2(P)}^2) \leq C(\mathcal{G}, \sigma, P)\kappa_n^2, \tag{103}$$

where $C(\mathcal{G}, \sigma, P) \in \mathbb{R}_+$ is a constant depending on $\mathcal{G}, \sigma$, and P only. In summary, the optimal sample complexity rate can be characterized by the sequence $(\kappa_n^2)_{n=1}^\infty$. We now particularize the Yang-Barron framework to $\mathcal{G} = H^1([0, 1])$ with P the uniform distribution on $[0, 1]$. First, note that

$$c_1\varepsilon^{-1} \leq \log(N(\varepsilon, H^1([0,1]), L^2([0,1]))) \leq \log(M(\varepsilon, H^1([0,1]), L^2([0,1]))) \leq C_1\varepsilon^{-1}, \ \varepsilon \in (0, \varepsilon_0), \tag{104}$$

where $\varepsilon_0, c_1, C_1 \in \mathbb{R}_+$ are absolute constants. Here, the first inequality follows from Lemma 3.5 and in the second inequality we used Lemma F.1. The last inequality in (104) is thanks to

$$
\begin{aligned}
\log(M(\varepsilon, H^1([0,1]), L^2([0,1]))) &\leq \log(N(\varepsilon/2, H^1([0,1]), L^2([0,1]))) \\
&\leq \log(N(\varepsilon/2, H^1([0,1]), L^\infty([0,1]))) \\
&\leq C_1\varepsilon^{-1},
\end{aligned}
$$

---

[4]Concretely, [35, Condition 2] requires that there exist an $\alpha \in (0, 1)$ such that $\liminf_{\varepsilon \to 0} \log(M(\alpha\varepsilon, \mathcal{G}, L^2(P)))/\log(M(\varepsilon, \mathcal{G}, L^2(P))) > 1$.

where in the first inequality we again used Lemma F.1, the second inequality follows from the fact that coverings with respect to the $L^\infty([0,1])$-norm are also coverings with respect to the $L^2([0,1])$-norm, and in the last inequality we used [6, Eq. 5.12]. Next, take $n \in \mathbb{N}$ large enough[5] for the solution κ_n of the equation $\kappa_n^2 = \frac{\log(M(\kappa_n, H^1([0,1]), L^2([0,1])))}{n}$ to satisfy $\kappa_n < \varepsilon_0$. We then have $c_1 \kappa_n^{-1} \leq \log(M(\kappa_n, H^1([0,1]), L^2([0,1]))) \leq C_1 \kappa_n^{-1}$, which implies $\frac{c_1 \kappa_n^{-1}}{n} \leq \kappa_n^2 \leq \frac{C_1 \kappa_n^{-1}}{n}$, and hence

$$c_1^{2/3} n^{-2/3} \leq \kappa_n^2 \leq C_1^{2/3} n^{-2/3}, \tag{105}$$

thereby recovering the optimal rate $n^{-2/3}$ mentioned in Corollary 4.2.

We proceed to derive sufficient conditions for a sequence of estimators to achieve optimal sample complexity rate. These conditions are general in the sense of the estimators not having to be neural networks and include, e.g., sparse dictionary approximation [36, 37, 38].

Corollary 4.3. *Let $\mathbb{X} \subseteq \mathbb{R}^d$ and consider the class $\mathcal{G}$ of regression functions mapping $\mathbb{X}$ to $\mathbb{R}$. Let $g \in \mathcal{G}$, $n \in \mathbb{N}$, and $\sigma \in \mathbb{R}_+$. Let P be a distribution on $\mathbb{X}$, with the associated samples $(x_i, y_i)_{i=1}^n = (x_i, g(x_i) + \sigma \xi_i)_{i=1}^n$, where $(x_i)_{i=1}^n$ are i.i.d. random variables of distribution P, $(\xi_i)_{i=1}^n$ are i.i.d. standard Gaussian random variables, and $(x_i)_{i=1}^n$ and $(\xi_i)_{i=1}^n$ are statistically independent.*

Let $\varepsilon_n \in (0, 1/2)$ and consider a class of functions $\mathcal{F}_n \subseteq L^\infty(\mathbb{X})$ such that

$$\mathcal{A}(\mathcal{G}, \mathcal{F}_n, \|\cdot\|_{L^2(P)}) \leq \varepsilon_n. \tag{106}$$

Let $\hat{f}_n$ be the empirical risk minimizer in $\mathcal{F}_n$, i.e.,

$$\frac{1}{n} \sum_{i=1}^n (\hat{f}_n(x_i) - y_i)^2 = \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 \right), \quad a.s. \tag{107}$$

Then,

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \leq C(1 + \sigma + \sigma^2 + (R(\mathcal{G}, \mathcal{F}_n))^2)\left(\frac{\varepsilon_n^2}{\kappa_n^2} + K(\mathcal{G}, \mathcal{F}_n, \varepsilon_n, \kappa_n, P) \right)\kappa_n^2, \tag{108}$$

where $C \in \mathbb{R}_+$ is an absolute constant, $R(\mathcal{G}, \mathcal{F}_n) := \max\{\sup_{h \in \mathcal{G}} \|h\|_{L^\infty(\mathbb{X})}, \sup_{f \in \mathcal{F}_n} \|f\|_{L^\infty(\mathbb{X})}\}$, κ_n is the solution to (102), and

$$K(\mathcal{G}, \mathcal{F}_n, \varepsilon_n, \kappa_n, P) = \frac{\log(N(\varepsilon_n^2, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{\log(M(\kappa_n, \mathcal{G}, L^2(P)))}. \tag{109}$$

Proof. We have

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \tag{110}$$

$$\leq C(1 + \sigma + \sigma^2 + (R(g, \mathcal{F}_n))^2)\left(\varepsilon_n^2 + \frac{\log(N(\varepsilon_n^2, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n} \right) \tag{111}$$

$$\leq C(1 + \sigma + \sigma^2 + (R(\mathcal{G}, \mathcal{F}_n))^2)\left(\varepsilon_n^2 + \frac{\log(N(\varepsilon_n^2, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n} \right) \tag{112}$$

$$= C(1 + \sigma + \sigma^2 + (R(\mathcal{G}, \mathcal{F}_n))^2)\left(\varepsilon_n^2 + K(\mathcal{G}, \mathcal{F}_n, \varepsilon_n, \kappa_n, P) \cdot \frac{\log(M(\kappa_n, \mathcal{G}, L^2(P)))}{n} \right) \tag{113}$$

$$= C(1 + \sigma + \sigma^2 + (R(\mathcal{G}, \mathcal{F}_n))^2)\left(\frac{\varepsilon_n^2}{\kappa_n^2} + K(\mathcal{G}, \mathcal{F}_n, \varepsilon_n, \kappa_n, P) \right)\kappa_n^2, \tag{114}$$

---

[5]It suffices to take $n \geq \lceil \frac{8C_1}{\varepsilon_0^3} \rceil + 1$, which we show leads to $\kappa_n < \frac{\varepsilon_0}{2}$. Suppose, for the sake of contradiction, that for such n, it holds that $\kappa_n \geq \frac{\varepsilon_0}{2}$. It would then follow from the monotonicity of the packing number that $\log(M(\kappa_n, H^1([0,1]), L^2([0,1]))) \leq \log(M(\frac{\varepsilon_0}{2}, H^1([0,1]), L^2([0,1]))) \leq 2C_1 \varepsilon_0^{-1}$, which together with $\kappa_n^2 = \frac{\log(M(\kappa_n, H^1([0,1]), L^2([0,1])))}{n}$ implies $\kappa_n^2 \leq \frac{2C_1 \varepsilon_0^{-1}}{n} \leq \frac{2C_1 \varepsilon_0^{-1}}{\lceil 8C_1/\varepsilon_0^3 \rceil + 1} < \frac{\varepsilon_0^2}{4}$. Hence, $\kappa_n < \frac{\varepsilon_0}{2}$, which establishes the contradiction.

where (111) follows from Theorem 4.1 with $\varepsilon = \varepsilon_n$ and the prerequisites (78) and (79) satisfied thanks to (106) and (107), respectively, in (112) we used $R(g, \mathcal{F}_n) \leq R(\mathcal{G}, \mathcal{F}_n)$, and (114) is by (102). $\qquad\square$

Assumption (106) in Corollary 4.3 is similar to the approximation assumption (78) in Theorem 4.1, only here we need to control the worst-case error over the entire class $\mathcal{G}$ of regression functions while Theorem 4.1 pertains to a fixed $g \in \mathcal{G}$. Assumption (107) corresponds to the empirical risk minimization condition (79) in Theorem 4.1, with the qualification that here we require exact empirical risk minimization while Theorem 4.1 allows for an additive slack term, given by ε^2.

Thanks to the Yang-Barron lower bound (101), we can now conclude from Corollary 4.3, specifically from (108), that for a sequence $\hat{f}_n$, $n \in \mathbb{N}$, of empirical risk minimizers to be optimal (up to constant factors), it suffices to meet the conditions (106) and (107) and have the quantity

$$C(1 + \sigma + \sigma^2 + (R(\mathcal{G}, \mathcal{F}_n))^2)\left(\frac{\varepsilon_n^2}{\kappa_n^2} + K(\mathcal{G}, \mathcal{F}_n, \varepsilon_n, \kappa_n, P)\right) \tag{115}$$

be upper-bounded by a constant not depending on n.

We next illustrate how the choices made in Corollary 4.2, specifically for $L(n)$ and $\mathcal{F}_n$ in (82), meet all these conditions thereby proving that the estimation of 1-Lipschitz functions can be accomplished through very deep ReLU networks in an information-theoretically optimal manner. Accordingly, we take $\mathcal{G} = H^1([0,1])$, let $L(n)$ and $\mathcal{F}_n$ be as in (82), and, for concreteness, take P to be the uniform distribution on $[0,1]$. First, we verify that the assumptions in Corollary 4.2 imply (106) and (107). Condition (107) is identical to (83) in Corollary 4.2. The approximation assumption (106) is satisfied with

$$\varepsilon_n = \frac{1}{4}n^{-1/3} \tag{116}$$

as (86)-(91) holds for all $g \in H^1([0,1])$.

We proceed to upper-bound the individual terms in (115) and start by noting that

$$R(\mathcal{G}, \mathcal{F}_n) = R(H^1([0,1]), \mathcal{F}_n) = \max\left\{\sup_{h \in H^1([0,1])} \|h\|_{L^\infty(\mathbb{X})}, \sup_{f \in \mathcal{F}_n} \|f\|_{L^\infty(\mathbb{X})}\right\} \leq 1, \tag{117}$$

thanks to the truncation operation $\mathcal{T}_1$ in the definition of $\mathcal{F}_n$. Second, we need to verify that $\mathcal{F}_n$ is such that ε_n^2 is balanced with κ_n^2 in the sense of $\frac{\varepsilon_n^2}{\kappa_n^2}$ being upper-bounded by a constant independent of n. This follows from

$$\frac{\varepsilon_n^2}{\kappa_n^2} \leq \frac{\frac{1}{16}n^{-2/3}}{c_1^{2/3}n^{-2/3}} = \frac{1}{16c_1^{2/3}}, \tag{118}$$

where we used (105) and (116).

To upper-bound $K(H^1([0,1]), \mathcal{F}_n, \varepsilon_n, \kappa_n, P)$, we first factorize according to

$$K(H^1([0,1]), \mathcal{F}_n, \varepsilon_n, \kappa_n, P) = \frac{\log(N(\varepsilon_n^2, \mathcal{F}_n, L^\infty([0,1]))) + 1}{\log(M(\kappa_n, H^1([0,1]), L^2([0,1])))} \tag{119}$$

$$= \frac{\log(N(\varepsilon_n^2, \mathcal{F}_n, L^\infty([0,1]))) + 1}{\log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^2([0,1])))}$$

$$\cdot \frac{\log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^2([0,1])))}{\log(N(4\varepsilon_n, H^1([0,1]), L^2([0,1])))}$$

$$\cdot \frac{\log(N(4\varepsilon_n, H^1([0,1]), L^2([0,1])))}{\log(M(\kappa_n, H^1([0,1]), L^2([0,1])))} \tag{120}$$

and then treat the three factors in (120) individually.

(i) For the numerator of the first factor in (120), we have

$$\log(N(\varepsilon_n^2, \mathcal{F}_n, L^\infty([0,1]))) + 1 \tag{121}$$

$$= \log(N(\varepsilon_n^2, \mathcal{T}_1 \circ \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^\infty([0,1]))) + 1 \tag{122}$$

$$\leq \log(N(\varepsilon_n^2, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^\infty([0,1]))) + 1 \tag{123}$$

$$\leq C_2 \lceil D+1 \rceil^2 L(n) \log\left(\frac{(\lceil D+1 \rceil + 1)^{L(n)}}{\varepsilon_n^2}\right) + 1 \tag{124}$$

$$\leq (C_2 + 1) \lceil D+1 \rceil^2 L(n) \log\left(\frac{(\lceil D+1 \rceil + 1)^{L(n)}}{\varepsilon_n^2}\right), \tag{125}$$

where (123) follows from the same argument as used to arrive at (94), and in (124) we employed the covering number upper bound in Theorem 2.1. For the denominator $\log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^2([0,1])))$, we apply the covering number lower bound[6] in Theorem 2.1, resulting in

$$\log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^2([0,1]))) \tag{126}$$

$$\geq c_2 \lceil D+1 \rceil^2 L(n) \log\left(\frac{(\lceil D+1 \rceil + 1)^{L(n)}}{\varepsilon_n}\right). \tag{127}$$

Dividing (125) by (127), it follows that

$$\frac{C_2 + 1}{c_2} \log\left(\frac{(\lceil D+1 \rceil + 1)^{L(n)}}{\varepsilon_n^2}\right) \bigg/ \log\left(\frac{(\lceil D+1 \rceil + 1)^{L(n)}}{\varepsilon_n}\right) \leq 2\frac{C_2 + 1}{c_2}, \tag{128}$$

where we used $(\lceil D+1 \rceil + 1)^{L(n)} < (\lceil D+1 \rceil + 1)^{2L(n)}$.

(ii) For the second factor in (120), the numerator can be upper-bounded according to

$$\log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^2([0,1]))) \tag{129}$$

$$\leq \log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^\infty([0,1]))) \tag{130}$$

$$\leq \log\left(N\left(\frac{1}{16}n^{-2/3}, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^\infty([0,1])\right)\right) \tag{131}$$

$$\leq C_5 n^{1/3}, \tag{132}$$

where (130) follows from the fact that coverings with respect to the $L^\infty([0,1])$-norm are also coverings with respect to the $L^2([0,1])$-norm, (131) is by $\frac{1}{16}n^{-2/3} < \varepsilon_n$, and (132) follows from (95)-(100). The denominator in the second factor in (120) can be lower-bounded by (104) according to

$$\log(N(4\varepsilon_n, H^1([0,1]), L^2([0,1]))) \geq c_1(4\varepsilon_n)^{-1} = c_1 n^{1/3}. \tag{133}$$

Using (132) and (133), we finally obtain

$$\frac{\log(N(\varepsilon_n, \mathcal{R}(1, \lceil D+1 \rceil, L(n), 1), L^2([0,1])))}{\log(N(4\varepsilon_n, H^1([0,1]), L^2([0,1])))} \leq \frac{C_5 n^{1/3}}{c_1 n^{1/3}} \leq \frac{C_5}{c_1}. \tag{134}$$

---

[6]The application of the lower bound requires $W, L \geq 60$, which holds thanks to $L(n) = \lceil 2(D+1)(C+1)^{1/2}n^{1/6} \rceil$, $W = \lceil D+1 \rceil$, with D the constant specified in Lemma 3.4 satisfying $D \geq 60$ owing to the specifics in the proof of [18, Theorem 3.1].

(iii) For the third factor in (120), we have

$$\frac{\log(N(4\varepsilon_n, H^1([0,1]), L^2([0,1])))}{\log(M(\kappa_n, H^1([0,1]), L^2([0,1])))} \leq \frac{C_1(4\varepsilon_n)^{-1}}{c_1\kappa_n^{-1}} \leq \frac{C_1^{4/3}}{c_1}, \tag{135}$$

where the first inequality follows from (104) and in the second inequality we used (105). Note that here we exploited the fact that ε_n is of the same order as κ_n, namely $n^{-1/3}$.

Putting (i)-(iii) together, we have shown that (115) can, indeed, be upper-bounded by a constant not depending on n, thereby establishing the information-theoretic optimality of the sequence of neural network estimators $\hat{f}_n$ in Corollary 4.2. This was accomplished by exploiting three key properties. The first one is the $\log(1/\varepsilon_n)$-scaling behavior of the metric entropy of the set of approximants $\mathcal{R}(1, \lceil D+1 \rceil, L(n), 1)$ used to establish (121)-(128). Specifically, this scaling behavior is crucial in the last step (128). The second property is the ε^{-1}-scaling behavior of the metric entropy of the set of regression functions $H^1([0,1])$ used to arrive at (135). Such a scaling behavior is common for unit balls in function spaces, see, e.g., [10, Table 1]. The third property, leading to (134), states that the metric entropy of the set of regression functions $H^1([0,1])$ has to be balanced with that of the set of approximants $\mathcal{R}(1, \lceil D+1 \rceil, L(n), 1)$. We note that this balancing property can be relaxed to

$$\frac{\log(N(\varepsilon, \mathcal{F}_\varepsilon, L^2(\mathbb{X})))}{\log(N(4\varepsilon, \mathcal{G}, L^2(\mathbb{X})))} \leq r_\mathcal{G}(\log(N(4\varepsilon, \mathcal{G}, L^2(\mathbb{X})))), \tag{136}$$

where $\mathcal{G}$ is the class of regression functions under consideration, $\mathcal{F}_\varepsilon$ denotes a set of approximants satisfying $\mathcal{A}(\mathcal{G}, \mathcal{F}_\varepsilon, \|\cdot\|) \leq \varepsilon$, and $r_\mathcal{G} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is such that $\limsup_{x\to+\infty} \frac{\log(r_\mathcal{G}(x))}{\log(x)} = 0$. In this case the upper bound (108) in Corollary 4.3 would still guarantee optimal sample complexity rate, but would exhibit an additional logarithmic factor. An example of such a behavior, albeit in the context of optimal approximation (through neural networks) rather than regression, can be found in [10] where it is referred to as Kolmogorov-Donoho optimal approximation, defined through $\mathcal{A}(\mathcal{G}, \mathcal{F}_\varepsilon, \|\cdot\|_{L^2(\mathbb{X})}) \leq \varepsilon$ and (136) holding concurrently. Note that $\mathcal{A}(\mathcal{G}, \mathcal{F}_\varepsilon, \|\cdot\|_{L^2(\mathbb{X})}) \leq \varepsilon$ implies $\frac{\log(N(\varepsilon, \mathcal{F}_\varepsilon, L^2(\mathbb{X})))}{\log(N(4\varepsilon, \mathcal{G}, L^2(\mathbb{X})))} \geq 1$. Although not explicitly mentioned and formally established, Kolmogorov-Donoho optimal neural network approximation is what leads to optimal sample complexity rates up to a logarithmic factor in [13, 14, 21]. We remark that Kolmogorov-Donoho optimality can also be achieved through sparse dictionary approximation [36, 37, 38]. Thanks to its generality, Corollary 4.3 allows to conclude that taking the set of approximants $\mathcal{F}_n$ to be obtained by sparse dictionary approximation, sample complexity rate-optimal regression of $\mathcal{G}$ up to a logarithmic factor is guaranteed whenever $\mathcal{F}_n$ achieves Kolmogorov-Donoho optimal approximation of $\mathcal{G}$.

5 Sparse Networks with Uniformly Bounded Weights

Sparse neural networks exhibit a connectivity s that is (typically much) smaller than the total number of weights $L(W^2 + W)$ in the network. In practical applications sparsity is often enforced with the goal of minimizing the amount of memory needed to store the network. The approximation-theoretic limits of sparse neural networks have been studied widely in the literature, starting with [22, 23, 25] and further considered, both in the context of function approximation and regression, in [10, 13, 14, 21, 24]. In the spirit of Theorem 2.1, we next characterize the covering numbers of sparse ReLU networks with uniformly bounded weights.

Theorem 5.1. *Let $p \in [1, \infty]$, $d, W, L, s \in \mathbb{N}$, $B, \varepsilon \in \mathbb{R}_+$, with $s \geq \max\{W, L\}$ and $B \geq 1$. Then, for all $\varepsilon \in (0, 1/2)$, we have*

$$\log(N(\varepsilon, \mathcal{R}(d, W, L, B, s), L^p([0,1]^d))) \leq C \min\{s, W^2 L\} \log\left(\frac{(W+1)^L B^L}{\varepsilon}\right), \qquad (137)$$

where $C \in \mathbb{R}_+$ is an absolute constant. Moreover, there exist absolute constants $c, D \in \mathbb{R}_+$ such that, if, in addition, $W, L \geq 60$ and $s \geq D d^2 L$, then, for all $\varepsilon \in (0, 1/4)$, it holds that

$$\log(N(\varepsilon, \mathcal{R}(d, W, L, B, s), L^p([0,1]^d))) \geq c \min\{s, W^2 L\} \log\left(\frac{(\widetilde{W}+1)^L B^L}{\varepsilon}\right), \qquad (138)$$

where $\widetilde{W} := \min\{\lceil \sqrt{\frac{s}{L}} \rceil, W\}$.

Proof. The proof is provided in Appendix D. $\qquad\square$

We first note that the condition $s \geq \max\{W, L\}$ comes without loss of generality, for the following reasons. A network with $s < L$ necessarily has one or more layers with weights equal to 0 and hence realizes a constant function, which could equivalently be obtained by a single-layer network. Likewise, for $s < W$, there would be nodes that can be removed without affecting the network's input-output relation.

We proceed to discuss the effect of the connectivity parameter s on the covering number bounds (137) and (138). First, recall that networks in $\mathcal{N}(d, W, L, B)$ have no more than $L(W^2 + W) \leq 2W^2 L$ weights. Comparing the factors in front of the logarithms in (137) and (138) to those in the corresponding bounds (3) and (4) for the fully-connected case, hence suggests an interpretation of $\min\{s, W^2 L\}$ as the effective connectivity. An important difference between the bounds for the fully-connected case in Theorem 2.1 and those in Theorem 5.1 is the appearance of the quantity $\widetilde{W} = \min\{\lceil \sqrt{\frac{s}{L}} \rceil, W\}$ inside the logarithm in the lower bound (138). For $\lceil \sqrt{\frac{s}{L}} \rceil < W$, we will hence have a loss of tightness, albeit only of logarithmic order, between the bounds (137) and (138). The term $\frac{s}{L}$ can be interpreted as the average connectivity per layer, a quantity also appearing in the VC-dimension lower bound [9, Equation (2)] for ReLU networks.

We finally note that the fundamental limits of sparse ReLU networks when used in neural network transformation, function approximation, and optimal regression, can be inferred by following the playbooks in Sections 3 and 4, but with the covering number behavior as quantified by Theorem 5.1.

6 Fully-connected Networks with Base-2 Quantized Weights

In this section, we characterize the covering number of ReLU networks with base-2 quantized weights, i.e., we consider the set $\mathcal{R}_{\mathbb{Q}_b^a}(d, W, L)$ with $\mathbb{Q}_b^a := (-2^{a+1}, 2^{a+1}) \cap 2^{-b}\mathbb{Z}$, where $a, b \in \mathbb{N}$. The motivation for analyzing this setting stems from the fact that neural networks stored on electronic devices necessarily have their weights encoded into finite-length bitstrings. For ease of presentation, we simplify notation according to

$$\begin{aligned}
\mathcal{N}_b^a(d, W, L) &:= \mathcal{N}_{\mathbb{Q}_b^a}(d, W, L) \\
&= \{\Phi \in \mathcal{N}(d) : \mathcal{W}(\Phi) \leq W, \ \mathcal{L}(\Phi) \leq L, \operatorname{coef}(\Phi) \subseteq \mathbb{Q}_b^a\}, \\
\mathcal{R}_b^a(d, W, L) &:= \{R(\Phi) : \Phi \in \mathcal{N}_b^a(d, W, L)\}.
\end{aligned}$$

To the best of our knowledge, there are no results in the literature on covering numbers of ReLU networks with base-2 quantized weights. Here, we report covering number lower and upper bounds that are tight.

Theorem 6.1. *Let $p \in [1, \infty]$, $d, W, L, a, b \in \mathbb{N}$. For all $\varepsilon \in (0, 1/2)$, it holds that*

$$\log(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d))) \leq CW^2 L \cdot \min\left\{(a+b), \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right)\right\}, \quad (139)$$

with $C \in \mathbb{R}_+$ an absolute constant. Moreover, there exist absolute constants $c, D, E \in \mathbb{R}_+$ such that, for $W, L \geq D$ with $L(a+b) \geq E \log(W)$, and all $\varepsilon \in (0, \frac{1}{100})$,

$$\log(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d))) \geq cW^2 L \cdot \min\left\{(a+b), \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right)\right\}. \quad (140)$$

Proof. We start by proving the upper bound. Arbitrarily fix $\varepsilon \in (0, 1/2)$. As $\mathcal{R}_b^a(d, W, L)$ is an ε-covering of itself, we have $N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d)) \leq |\mathcal{R}_b^a(d, W, L)|$ and hence

$$\log(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d))) \leq \log(|\mathcal{R}_b^a(d, W, L)|) \quad (141)$$
$$\leq \log(|\mathcal{N}_b^a(d, W, L)|) \quad (142)$$
$$\leq 5W^2 L \log(|\mathbb{Q}_b^a|) \quad (143)$$
$$< 10W^2 L(a+b), \quad (144)$$

where (143) follows from Lemma 2.4 with $\mathbb{A} = \mathbb{Q}_b^a$, and (144) is by $|\mathbb{Q}_b^a| = |(-2^{a+1}, 2^{a+1}) \cap 2^{-b}\mathbb{Z}| \leq 2^{a+2} \cdot 2^b \leq 2^{2(a+b)}$. Moreover, as $\mathcal{R}_b^a(d, W, L) \subseteq \mathcal{R}(d, W, L, 2^{a+1})$, it holds that

$$\mathcal{A}(\mathcal{R}_b^a(d, W, L), \mathcal{R}(d, W, L, 2^{a+1}), \|\cdot\|_{L^p([0,1]^d)}) = 0. \quad (145)$$

Application of Proposition 3.1 with $\mathcal{G} = \mathcal{R}_b^a(d, W, L)$, $\mathcal{F} = \mathcal{R}(d, W, L, 2^{a+1})$, $\delta = \|\cdot\|_{L^p([0,1]^d)}$, ε replaced by $\varepsilon/4$, and the prerequisite (39) satisfied thanks to (145), now yields

$$N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d)) \leq N(\varepsilon/4, \mathcal{R}(d, W, L, 2^{a+1}), L^p([0,1]^d)). \quad (146)$$

The logarithm of the right-hand-side of (146) can be upper-bounded according to

$$\log(N(\varepsilon/4, \mathcal{R}(d, W, L, 2^{a+1}), L^p([0,1]^d))) \leq C_1 W^2 L \log\left(\frac{(W+1)^L 2^{(a+1)L}}{\varepsilon/4}\right) \quad (147)$$
$$< 3C_1 W^2 L \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right), \quad (148)$$

where in (147) we applied Theorem 2.1, (148) follows from $\frac{(W+1)^L 2^{(a+1)L}}{\varepsilon/4} < \frac{(W+1)^L 2^{3aL}}{\varepsilon \cdot \varepsilon^2} < \left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right)^3$, and $C_1 \in \mathbb{R}_+$ is an absolute constant. Combining (146) and (147)-(148), establishes

$$\log(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d))) < 3C_1 W^2 L \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right). \quad (149)$$

Putting (141)-(144) and (149) together finally yields

$$\log(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d)))$$
$$< \min\left\{10W^2 L(a+b), 3C_1 W^2 L \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right)\right\}$$
$$< (10 + 3C_1)W^2 L \min\left\{(a+b), \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right)\right\}.$$

Upon setting $C = 10 + 3C_1$, this concludes the proof of the upper bound. The proof of the lower bound is lengthy and hence relegated to Appendix E. $\square$

First, we note that the upper bound (139) and the lower bound (140) are tight up to an absolute multiplicative constant. This allows us to conclude that the covering number of ReLU networks with base-2 quantized weights exhibits two regimes as a function of ε. Specifically, for $\varepsilon \geq \frac{(W+1)^L 2^{aL}}{2^{a+b}}$, the log-term in the bounds (139) and (140) is active, which renders them structurally identical to the bounds for networks with unquantized weights, as stated in Theorem 2.1. In this regime quantized neural networks, in terms of their covering numbers, hence behave like unquantized networks. On the other hand, for $\varepsilon < \frac{(W+1)^L 2^{aL}}{2^{a+b}}$, the covering number can be sandwiched by quantities that are independent of ε and solely determined by the parameters W, L, a, b according to $cW^2 L(a+b) \leq \log(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d))) \leq CW^2 L(a+b)$. In this regime, the covering ball radius ε is small enough to reveal the quantized nature of the network weights. In summary, we have a phase-transition behavior, in terms of ε, between a regime where $\mathcal{N}_b^a(d, W, L)$ behaves like networks with weights in $\mathbb{R}$ and a regime where the quantized nature of the weights limits the approximation capacity of $\mathcal{N}_b^a(d, W, L)$.

We finally note that the fundamental limits of ReLU networks with base-2 quantized weights when used in neural network transformation, function approximation, and optimal regression, can be inferred by following the playbooks in Sections 3 and 4, but with the covering number behavior as quantified by Theorem 6.1.

7 Fully-connected Networks with Truncated Output

Fully-connected ReLU networks with unconstrained weight magnitude are prevalent in the literature [7, 8, 11, 39]. As the covering number of the function class $\mathcal{R}(d, W, L)$ realized by such networks is infinite, their performance limits are typically characterized through the VC-dimension. It turns out, however, that when dealing with bounded functions such as the set $H^1([0,1])$, it suffices to consider ReLU networks with truncated outputs. This allows to develop a more refined picture, namely by arguing as follows. First, note that $\mathcal{R}(1, W, L, 1) \subseteq \mathcal{R}(1, W, L)$ together with Lemma 3.4, yields

$$\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L), \|\cdot\|_{L^p([0,1])}) \leq \mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^p([0,1])}) \tag{150}$$
$$\leq C(W^2 L^2 \log(W))^{-1}, \tag{151}$$

for $p \in [1, \infty]$, $W, L \in \mathbb{N}$, with $W, L \geq D$, where $C, D \in \mathbb{R}_+$ are the absolute constants specified in Lemma 3.4. To the best of our knowledge a corresponding lower bound is available only for $p = \infty$, namely [18, Proposition 2.11], [7, Theorem 2.3],

$$\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L), \|\cdot\|_{L^\infty([0,1])}) \geq c\,(W^2 L^2(\log(W) + \log(L)))^{-1},$$

with $c \in \mathbb{R}_+$ an absolute constant. The goal of this section is to derive a lower bound for $p = 2$. This will require, inter alia, an upper bound on the covering number of ReLU networks with unconstrained weight magnitude and truncated output. To the best of our knowledge, there are no results available in the literature on covering numbers of ReLU networks with unconstrained weight magnitude and truncated output.

We proceed as follows. First, note that

$$\mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L), \|\cdot\|_{L^2([0,1])}) \tag{152}$$
$$= \sup_{f \in H^1([0,1])} \inf_{g \in \mathcal{R}(1,W,L)} \|f - g\|_{L^2([0,1])} \tag{153}$$
$$\geq \sup_{f \in H^1([0,1])} \inf_{g \in \mathcal{R}(1,W,L)} \|\mathcal{T}_1 \circ f - \mathcal{T}_1 \circ g\|_{L^2([0,1])} \tag{154}$$
$$= \sup_{f \in H^1([0,1])} \inf_{g \in \mathcal{R}(1,W,L)} \|f - \mathcal{T}_1 \circ g\|_{L^2([0,1])} \tag{155}$$
$$= \mathcal{A}(H^1([0,1]), \mathcal{T}_1 \circ \mathcal{R}(1, W, L), \|\cdot\|_{L^2([0,1])}), \tag{156}$$

where (154) follows from the fact that $\mathcal{T}_1$ is 1-Lipschitz, and in (155) we used that f is uniformly bounded by 1 on $[0,1]$. Therefore, to lower-bound $\mathcal{A}(H^1([0,1]), \mathcal{R}(1,W,L), \|\cdot\|_{L^2([0,1])})$, it suffices to lower-bound $\mathcal{A}(H^1([0,1]), \mathcal{T}_1 \circ \mathcal{R}(1,W,L), \|\cdot\|_{L^2([0,1])})$, which will be effected through the technique developed to prove the minimax approximation error lower bound (44) in Corollary 3.3 combined with a new covering number upper bound for $\mathcal{T}_1 \circ \mathcal{R}(1,W,L)$. We note that all our arguments can be extended, with minor effort, to the approximation of function classes that are uniformly bounded by arbitrary constants $E \in \mathbb{R}_+$, but we will stick to $H^1([0,1])$ for brevity of exposition.

We first present the upper bound on the covering number of $\mathcal{T}_1 \circ \mathcal{R}(d,W,L)$ and then show how it can be used to lower-bound $\mathcal{A}(H^1([0,1]), \mathcal{T}_1 \circ \mathcal{R}(1,W,L), \|\cdot\|_{L^2([0,1])})$.

Theorem 7.1. *Let $d,W,L \in \mathbb{N}$, with $W,L \geq 2$, and let P be a distribution on $\mathbb{R}^d$. For all $\varepsilon \in (0,1/2)$, it holds that*

$$\log(N(\varepsilon, \mathcal{T}_1 \circ \mathcal{R}(d,W,L), L^2(P))) \leq CW^2L^2 \log(WL) \log(\varepsilon^{-1}),$$

with $C \in \mathbb{R}_+$ an absolute constant.

The proof of Theorem 7.1 is based on a relation in [40] between the covering number with respect to the $L^2(P)$-norm, for arbitrary distributions P, and the fat-shattering dimension of uniformly-bounded function classes combined with bounds on the fat-shattering dimension of ReLU networks [9]. We first prepare the technical ingredients of the proof and start by recalling the definition of fat-shattering dimension.

Definition 7.2. *[40] Let $\mathcal{X}$ be a set, $\mathcal{F}$ a class of functions from $\mathcal{X}$ to $\mathbb{R}$, and $\gamma \in \mathbb{R}_+$. The fat-shattering dimension of $\mathcal{F}$, written as $fat(\mathcal{F}, \gamma)$, is the largest $m \in \mathbb{N}$ for which there exists $(x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for every $(b_1, \ldots, b_m) \in \{0,1\}^m$, there is an $f \in \mathcal{F}$ so that, for all $i \in \{1, \ldots, m\}$,*

$$f(x_i) \begin{cases} \geq y_i + \gamma, & \text{if } b_i = 1, \\ \leq y_i, & \text{if } b_i = 0. \end{cases} \tag{157}$$

Next, we upper-bound the fat-shattering dimension of $\mathcal{T}_1 \circ \mathcal{R}(d,W,L)$.

Lemma 7.3. *For $d,W,L \in \mathbb{N}$, with $W,L \geq 2$, it holds that*

$$fat(\mathcal{T}_1 \circ \mathcal{R}(d,W,L), \gamma) \leq C_h W^2 L^2 (\log(WL)), \quad \text{for all } \gamma \in \mathbb{R}_+, \tag{158}$$

with $C_h \in \mathbb{R}_+$ an absolute constant.

Proof. The result is essentially an implication of [9, Eq. (2)], with minor additional observations. We provide the details in Appendix B.4. $\square$

The Mendelson-Vershynin upper bound [40] on the covering number in terms of fat-shattering dimension is as follows.

Theorem 7.4. *[40, Theorem 1] Let $\mathcal{F}$ be a class of functions defined on a set $\mathcal{X}$ with $\sup_{x \in \mathcal{X}, f \in \mathcal{F}} |f(x)| \leq 1$. Then, for every distribution P on $\mathcal{X}$, and all $\varepsilon \in (0,1)$,*

$$N(\varepsilon, \mathcal{F}, L^2(P)) \leq \left(\frac{2}{\varepsilon}\right)^{K \cdot fat(\mathcal{F}, c\varepsilon)},$$

where $K, c \in \mathbb{R}_+$ are absolute constants.

We are now ready to prove Theorem 7.1.

Proof of Theorem 7.1. Application of Theorem 7.4 with $\mathcal{F} = \mathcal{T}_1 \circ \mathcal{R}(d, W, L)$, yields

$$N(\varepsilon, \mathcal{T}_1 \circ \mathcal{R}(d, W, L), L^2(P)) \leq \left(\frac{2}{\varepsilon}\right)^{K \cdot \text{fat}(\mathcal{T}_1 \circ \mathcal{R}(d,W,L), c\varepsilon)}. \tag{159}$$

Taking logarithms in (159) and applying Lemma 7.3 with $\gamma = c\varepsilon$, results in

$$\log(N(\varepsilon, \mathcal{T}_1 \circ \mathcal{R}(d, W, L), L^2(P))) \leq K \cdot C_h W^2 L^2 (\log(WL)) \log(2\varepsilon^{-1}).$$

The proof is concluded upon noting that $\log(2\varepsilon^{-1}) \leq \log(\varepsilon^{-2}) = 2\log(\varepsilon^{-1})$, for $\varepsilon \in (0, 1/2)$, and letting $C := 2K \cdot C_h$. □

We are now ready to put Theorem 7.1 to work in deriving the sought lower bound on $\mathcal{A}(H^1([0, 1]), \mathcal{R}(1, W, L), \| \cdot \|_{L^2([0,1])})$.

Corollary 7.5. *For $W, L \in \mathbb{N}$, with $W, L \geq 2$, it holds that*

$$\mathcal{A}(H^1([0, 1]), \mathcal{R}(1, W, L), \| \cdot \|_{L^2([0,1])}) \geq \min\left\{\frac{1}{8}, C(W^2 L^2 (\log(WL))^2)^{-1}\right\}, \tag{160}$$

with $C \in \mathbb{R}_+$ an absolute constant.

Proof. Set

$$\kappa := \mathcal{A}(H^1([0, 1]), \mathcal{R}(1, W, L), \| \cdot \|_{L^2([0,1])}). \tag{161}$$

For $\kappa \geq \frac{1}{8}$, (160) holds trivially. For $\kappa < \frac{1}{8}$, we first note that putting (161) together with (152)-(156), yields

$$\mathcal{A}(H^1([0, 1]), \mathcal{T}_1 \circ \mathcal{R}(1, W, L), \| \cdot \|_{L^2([0,1])}) \leq \kappa. \tag{162}$$

It then follows from Proposition 3.1 with $\varepsilon = \kappa$, $\mathcal{G} = H^1([0, 1])$, $\mathcal{F} = \mathcal{T}_1 \circ \mathcal{R}(1, W, L)$, $\delta = \| \cdot \|_{L^2([0,1])}$, and the prerequisite (39) satisfied thanks to (162), that

$$N(\kappa, \mathcal{T}_1 \circ \mathcal{R}(1, W, L), L^2([0, 1])) \geq N(4\kappa, H^1([0, 1]), L^2([0, 1])). \tag{163}$$

Next, we upper-bound the left-hand-side and lower-bound the right-hand-side of (163). For the former, we note that Theorem 7.1 with $d = 1$ and P the uniform distribution on $[0, 1]$, yields

$$\log(N(\kappa, \mathcal{T}_1 \circ \mathcal{R}(1, W, L), L^2([0, 1]))) \leq C_1 W^2 L^2 \log(WL) \log(\kappa^{-1}), \tag{164}$$

with $C \in \mathbb{R}_+$ an absolute constant. The lower bound on $N(4\kappa, H^1([0, 1]), L^2([0, 1]))$ is obtained from Lemma 3.5 with $\varepsilon = 4\kappa$ as

$$\log(N(4\kappa, H^1([0, 1]), L^p([0, 1]))) \geq C_2 (4\kappa)^{-1}, \tag{165}$$

with $C_2 \in \mathbb{R}_+$ an absolute constant. Combining (164) and (165) with (163), yields

$$C_1 W^2 L^2 \log(WL) \log(\kappa^{-1}) \geq C_2 (4\kappa)^{-1},$$

which implies

$$\frac{\kappa^{-1}}{\log(\kappa^{-1})} \leq C_3 W^2 L^2 \log(WL), \tag{166}$$

with $C_3 := \max\{\frac{4C_1}{C_2}, 4\} \in \mathbb{R}_+$ an absolute constant. We further upper-bound the right-hand-side of (166) according to

$$C_3 W^2 L^2 \log(WL) = \frac{8\log(C_3) C_3 W^2 L^2 (\log(WL))^2}{\log((WL)^{8\log(C_3)})} \tag{167}$$

$$\leq \frac{8\log(C_3) C_3 W^2 L^2 (\log(WL))^2}{\log(8\log(C_3) C_3 W^2 L^2 (\log(WL))^2)}, \tag{168}$$

27

where in (168) we used

$$\begin{aligned}
(WL)^{8\log(C_3)} &\geq (WL)^{4\log(C_3)} \cdot (WL)^4 \\
&\geq (C_3)^4 \cdot W^2 L^2 (\log(WL))^2 \\
&\geq 8\log(C_3) C_3 W^2 L^2 (\log(WL))^2.
\end{aligned}$$

Next, define $f : \mathbb{R}_+ \mapsto \mathbb{R}$ according to $f(x) = \frac{x}{\log(x)}$. Then, (166)-(168) can be written as

$$f(\kappa^{-1}) \leq f(8\log(C_3) C_3 W^2 L^2 (\log(WL))^2). \tag{169}$$

We note that $\kappa^{-1} > (\frac{1}{8})^{-1} > e$ and $8\log(C_3) C_3 W^2 L^2 (\log(WL))^2 > e$, and the function f is strictly increasing on (e, ∞) as $f'(x) = \ln(2)\frac{\ln(x)-1}{(\ln(x))^2} > 0$, for $x \in (e, \infty)$. It hence follows from (169) that $\kappa^{-1} \leq 8\log(C_3) C_3 W^2 L^2 (\log(WL))^2$, which is

$$\kappa = \mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L), \|\cdot\|_{L^2([0,1])}) \geq (8\log(C_3)C_3)^{-1}(W^2 L^2 (\log(WL))^2)^{-1}.$$

The proof is concluded upon taking $C = (8\log(C_3)C_3)^{-1}$. $\qquad\square$

As a byproduct of the results obtained in this section, we can conclude that in the approximation (in $L^2([0,1])$-norm) of functions in $H^1([0,1])$ going from networks with bounded weights to networks with unbounded weights does not substantially improve approximation accuracy. Specifically, we have the following chain of inequalities

$$\min\left\{\frac{1}{8}, C(W^2 L^2 (\log(WL))^2)^{-1}\right\} \leq \mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L), \|\cdot\|_{L^2([0,1])}) \tag{170}$$

$$\leq \mathcal{A}(H^1([0,1]), \mathcal{R}(1, W, L, 1), \|\cdot\|_{L^2([0,1])}) \tag{171}$$

$$\leq C_2(W^2 L^2 \log(W))^{-1}, \tag{172}$$

where (170) is the lower bound (160), (171) follows from $\mathcal{R}(1, W, L, 1) \subseteq \mathcal{R}(1, W, L)$, and (172) is Lemma 3.4 with $p = 2$. Here, $C, C_2 \in \mathbb{R}_+$ are absolute constants. This shows that the improvement obtainable from allowing unbounded weights is at most of order $\frac{\log(W)}{(\log(WL))^2}$.

A Notation and Basic Definitions

We denote the cardinality of the set X by $|X|$. $\mathbb{N} = \{1, 2, \dots\}$ designates the natural numbers, $\mathbb{R}$ stands for the real numbers, $\mathbb{R}_+$ for the positive real numbers, and $\emptyset$ for the empty set. The maximum, minimum, supremum, and infimum of the set $\mathbb{A} \subseteq \mathbb{R}$ are denoted by $\max \mathbb{A}$, $\min \mathbb{A}$, $\sup \mathbb{A}$, and $\inf \mathbb{A}$, respectively. The indicator function 1_P for proposition P is equal to 1 if P is true and 0 else. For a metric space $(\mathcal{X}, \delta)$ and sets $\mathcal{F}, \mathcal{G} \subseteq \mathcal{X}$, we define $\mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) = \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \delta(f, g)$. For a vector $b \in \mathbb{R}^d$, we let $\|b\|_\infty := \max_{i=1,\dots,d} |b_i|$, $\|b\|_0 := \sum_{i=1}^d 1_{b_i \neq 0}$, and $\|b\|_1 := \sum_{i=1}^d |b_i|$. Similarly, for a matrix $A \in \mathbb{R}^{m \times n}$, we define $\|A\|_\infty = \max_{i=1,\dots,m, j=1,\dots,n} |A_{i,j}|$ and $\|A\|_0 = \sum_{i=1}^m \sum_{j=1}^n 1_{A_{i,j} \neq 0}$. 1_m and 0_m stand for the m-dimensional vector with all entries equal to 1 and 0, respectively. I_m refers to the $m \times m$ identity matrix. $1_{m \times n}$ and $0_{m \times n}$ denote the $m \times n$ matrix with all entries equal to 1 and 0, respectively. We designate the block-diagonal matrix with diagonal element-matrices $A_1, \dots, A_n$, possibly of different dimensions, by $\text{diag}(A_1, \dots, A_n)$. The truncation operator $\mathcal{T}_E : \mathbb{R} \mapsto [-E, E]$, $E \in \mathbb{R}_+$, is $\mathcal{T}_E(x) = \max\{-E, \min\{E, x\}\}$. $\log(\cdot)$ and $\ln(\cdot)$ denote the logarithm to base 2 and base e, respectively. The ReLU activation function is defined as $\rho(x) = \max\{x, 0\}$, for $x \in \mathbb{R}$, and, when applied to vectors, acts elementwise. The sign function $\text{sgn} : \mathbb{R} \mapsto \{0, 1\}$ is given by $\text{sgn}(x) = 1$, for $x \geq 0$, and $\text{sgn}(x) = 0$, for $x < 0$. We use $S(A, b)$ to refer to the affine mapping

$S(A, b)(x) = Ax + b, x \in \mathbb{R}^{n_2}$, with $A \in \mathbb{R}^{n_1 \times n_2}$, $b \in \mathbb{R}^{n_1}$. For the set $\mathbb{X} \subseteq \mathbb{R}^d$, with $d \in \mathbb{N}$, and the function $f : \mathbb{X} \mapsto \mathbb{R}$, we define the $L^p(\mathbb{X})$-norm of f, with $p \in [1, \infty)$, according to $\|f\|_{L^p(\mathbb{X})} = (\int_{x \in \mathbb{X}} |f(x)|^p d\mu(x))^{1/p}$, where μ is the Lebesgue measure on $\mathbb{R}^d$. The $L^\infty(\mathbb{X})$-norm of f is given by $\|f\|_{L^\infty(\mathbb{X})} = \sup_{x \in \mathbb{X}} |f(x)|$, and, for a distribution P on $\mathbb{X}$, we define the $L^2(P)$-norm of f as $\|f\|_{L^2(P)} = (\int_{x \in \mathbb{X}} |f(x)|^2 dP(x))^{1/2}$. A constant is said to be absolute if it does not depend on any variables or parameters.

B Further Proofs

B.1 Proof of Proposition 3.1

We start with a lemma that gives a lower bound on the cardinality of $\mathcal{F}$, in terms of the packing number of $\mathcal{G}$, under the condition (39).

Lemma B.1. *Let $(\mathcal{X}, \delta)$ be a metric space, $\mathcal{F}, \mathcal{G} \subseteq \mathcal{X}$, and $\varepsilon \in \mathbb{R}_+$. Assume that $\mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) \leq \varepsilon$. Then, we have*

$$|\mathcal{F}| \geq M(2\varepsilon, \mathcal{G}, \delta).$$

Proof. Arbitrarily fix $\varepsilon \in \mathbb{R}_+$. Suppose, for the sake of contradiction, that $|\mathcal{F}| < M(2\varepsilon, \mathcal{G}, \delta)$, which would imply the existence of a (2ε)-packing $\mathcal{P}$ of $\mathcal{G}$ such that $|\mathcal{F}| < |\mathcal{P}|$. In particular, $\mathcal{F}$ would be a finite set. Since $\mathcal{P}$ is a subset of $\mathcal{G}$, we have $\mathcal{A}(\mathcal{P}, \mathcal{F}, \delta) \leq \mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) \leq \varepsilon$, and therefore every element of $\mathcal{P}$ would be contained in an ε-neighborhood, with respect to the metric δ, of an $f \in \mathcal{F}$. As $|\mathcal{F}| < |\mathcal{P}|$ and there are $|\mathcal{F}|$ such neighborhoods and $|\mathcal{P}|$ elements to be contained in neighborhoods, the pigeonhole principle implies the existence of $x_1, x_2 \in \mathcal{P}$ such that $x_1, x_2 \in \{x : \delta(f_0, x) \leq \varepsilon\}$ for some $f_0 \in \mathcal{F}$. It would then follow from the triangle inequality that $\delta(x_1, x_2) \leq \delta(x_1, f_0) + \delta(f_0, x_2) \leq 2\varepsilon$, which implies that $\mathcal{P}$ can not be a (2ε)-packing. This establishes the desired contradiction. $\qquad\square$

We are now ready to prove Proposition 3.1.

If $N(\varepsilon, \mathcal{F}, \delta) = \infty$, then (40) holds trivially. For $N(\varepsilon, \mathcal{F}, \delta) < \infty$, suppose that $\mathcal{C}$ is a minimal ε-covering of $\mathcal{F}$. Defining $p : \mathcal{F} \to \mathcal{C}$ according to $p(f) = \operatorname{argmin}_{c \in \mathcal{C}} \delta(f, c)$, we hence get

$$\delta(f, p(f)) \leq \varepsilon. \tag{173}$$

Elements of $\mathcal{G}$ can now be approximated by elements of $\mathcal{C}$, with corresponding minimax approximation error

$$\mathcal{A}(\mathcal{G}, \mathcal{C}, \delta) = \sup_{g \in \mathcal{G}} \inf_{c \in \mathcal{C}} \delta(g, c) \tag{174}$$

$$= \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \inf_{c \in \mathcal{C}} \delta(g, c) \tag{175}$$

$$\leq \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \inf_{c \in \mathcal{C}} (\delta(g, f) + \delta(f, c)) \tag{176}$$

$$= \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} (\delta(g, f) + \delta(f, p(f))) \tag{177}$$

$$\leq \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} (\delta(g, f) + \varepsilon) \tag{178}$$

$$= \mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) + \varepsilon, \tag{179}$$

$$\leq 2\varepsilon. \tag{180}$$

where (176) is by the triangle inequality, (177) follows by definition of p, in (178) we used (173), and (180) is thanks to the assumption $\mathcal{A}(\mathcal{G}, \mathcal{F}, \delta) \leq \varepsilon$. Application of Lemma B.1 with $\mathcal{F} = \mathcal{C}$, ε replaced by 2ε, and the prerequisite satisfied thanks to (174)-(180), yields

$$N(\varepsilon, \mathcal{F}, \delta) = |\mathcal{C}| \geq M(4\varepsilon, \mathcal{G}, \delta),$$

which together with $M(4\varepsilon, \mathcal{G}, \delta) \geq N(4\varepsilon, \mathcal{G}, \delta)$, owing to Lemma F.1, concludes the proof.

B.2 Proof of Lemma 2.5

Fix a maximal (2ε)-packing $\{f_i\}_{i=1}^{M(2\varepsilon, \mathcal{R}(1,W,L,B), L^1([0,1]))}$ of $\mathcal{R}(1, W, L, B)$ with respect to the $L^1([0,1])$-norm. We shall lift this packing into a (2ε)-packing of $\mathcal{R}(d, W, L, B)$ with respect to the $L^p([0,1]^d)$-norm. Specifically, for $i \in \{1, \ldots, M(2\varepsilon, \mathcal{R}(1, W, L, B), L^1([0,1]))\}$, as $f_i \in \mathcal{R}(1, W, L, B)$, there exists a network configuration $\Phi_i = (A_\ell^i, b_\ell^i)_{\ell=1}^{\tilde{L}_i} \in \mathcal{N}(1, W, L, B)$ with $\tilde{L}_i \leq L$ such that $R(\Phi_i) = f_i$. Let $(\tilde{A}_1^i, \tilde{b}_1^i) = ((A_1^i, 0_{d' \times (d-1)}), b_1^i)$, with d' the number of rows of A_1^i, and $(\tilde{A}_\ell^i, \tilde{b}_\ell^i) = (A_\ell^i, b_\ell^i)$, for $1 < \ell \leq \tilde{L}_i$, and set $g_i := R((\tilde{A}_\ell^i, \tilde{b}_\ell^i)_{\ell=1}^{\tilde{L}_i})$. We note that, for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$,
$$\tilde{A}_1^i(x_1, \ldots, x_d)^T + \tilde{b}_1^i = A_1^i x_1 + b_1^i,$$
which implies, for all $(x_1, \ldots, x_d) \in \mathbb{R}^d$,
$$g_i(x_1, \ldots, x_d) = R((\tilde{A}_\ell^i, \tilde{b}_\ell^i)_{\ell=1}^{\tilde{L}_i})(x_1, \ldots, x_d) = R((A_\ell^i, b_\ell^i)_{\ell=1}^{\tilde{L}_i})(x_1) = f_i(x_i). \tag{181}$$
As $(\tilde{A}_\ell^i, \tilde{b}_\ell^i)_{\ell=1}^{\tilde{L}_i} \in \mathcal{N}(d, W, L, B)$, we have $g_i \in R(d, W, L, B)$. Next, we shall establish that $\{g_i\}_{i=1}^{M(2\varepsilon, \mathcal{R}(1,W,L,B), L^1([0,1]))}$ is a (2ε)-packing of $\mathcal{R}(d, W, L, B)$ with respect to the $L^p([0,1]^d)$-norm. To this end, let $q \in [1, \infty]$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. We then have, for $i, j \in \{1, \ldots, M(2\varepsilon, \mathcal{R}(1, W, L, B), L^1([0,1]))\}$ with $i \neq j$, that

$$\|g_i - g_j\|_{L^p([0,1]^d)} = \|g_i - g_j\|_{L^p([0,1]^d)} \|\mathbf{1}\|_{L^q([0,1]^d)} \tag{182}$$
$$\geq \|g_i - g_j\|_{L^1([0,1]^d)} \tag{183}$$
$$= \int_{(x_1, \ldots, x_d) \in [0,1]^d} |g_i(x_1, \ldots, x_d) - g_j(x_1, \ldots, x_d)| \, dx_1 \ldots dx_d \tag{184}$$
$$= \int_{(x_1, \ldots, x_d) \in [0,1]^d} |f_i(x_1) - f_j(x_1)| \, dx_1 \ldots dx_d \tag{185}$$
$$= \int_{x_1 \in [0,1]} |f_i(x_1) - f_j(x_1)| dx_1 \tag{186}$$
$$= \|f_i - f_j\|_{L^1([0,1])} \tag{187}$$
$$> 2\varepsilon, \tag{188}$$

where in (182) we denoted by $\mathbf{1}$ the constant function taking value 1 on $[0,1]^d$, (183) follows from the Hölder inequality, in (185) we used (181), and (188) is a consequence of $\{f_i\}_{i=1}^{M(2\varepsilon, \mathcal{R}(1,W,L,B), L^1([0,1]))}$ being a (2ε)-packing with respect to the $L^1([0,1])$-norm. We have therefore established that $\{g_i\}_{i=1}^{M(2\varepsilon, \mathcal{R}(1,W,L,B), L^1([0,1]))}$ is a (2ε)-packing of $\mathcal{R}(d, W, L, B)$ with respect to the $L^p([0,1]^d)$-norm, and hence

$$M(2\varepsilon, \mathcal{R}(d, W, L, B), L^p([0,1]^d)) \geq M(2\varepsilon, \mathcal{R}(1, W, L, B), L^1([0,1])). \tag{189}$$

B.3 Proof of Lemma 2.7

For $\varepsilon \geq \frac{E}{4N}$, we have $\log(\lceil \frac{E}{4\varepsilon N} \rceil) = 0$ so that (23) holds trivially. For $\varepsilon < \frac{E}{4N}$, we prove the statement by explicitly constructing an ε-packing of suitable cardinality. To this end, for $y = (y_i)_{i=1}^N$, we define functions $f_y \in \Sigma(X_N, \infty)$ as follows

$$f_y(x) = \begin{cases} 0, & \text{for } x \in (-\infty, 0], \\ f_y(\frac{i-1}{N}) + N(x - \frac{i-1}{N})(y_i - f_y(\frac{i-1}{N})), & \text{for } x \in (\frac{i-1}{N}, \frac{i}{N}], \quad i = 1, \ldots, N, \\ y_N, & \text{for } x \in (1, \infty). \end{cases} \tag{190}$$

and note that $f_y(0) = 0$ and $f_y(\frac{i}{N}) = y_i$, for $i = 1, \ldots, N$. Now, consider the set of functions

$$\mathcal{F}_{N,M} := \left\{ f_y : y = (y_i)_{i=1}^N \in \left(\left\{ \frac{\ell}{M} E \right\}_{\ell=0}^M \right)^N \right\} \subseteq \Sigma(X_N, E),$$

with M an integer to be specified later, namely such that $\mathcal{F}_{N,M}$ is an ε-packing of $\Sigma(X_N, E)$ with appropriate cardinality. We proceed to derive a lower bound on the distance between distinct elements in $\mathcal{F}_{N,M}$. For $y = (y_i)_{i=1}^N \in (\{\frac{\ell}{M} E\}_{\ell=0}^M)^N$ and $\tilde{y} = (\tilde{y}_i)_{i=1}^N \in (\{\frac{\ell}{M} E\}_{\ell=0}^M)^N$ such that $y \neq \tilde{y}$, we let $j \in \{1, \ldots, N\}$ be the smallest index for which $y_j \neq \tilde{y}_j$, and then get

$$\|f_y - f_{\tilde{y}}\|_{L^1([0,1])} \tag{191}$$

$$= \int_0^1 |f_y(x) - f_{\tilde{y}}(x)| dx \tag{192}$$

$$\geq \int_{\frac{j-1}{N}}^{\frac{j}{N}} |f_y(x) - f_{\tilde{y}}(x)| dx \tag{193}$$

$$= \int_{\frac{j-1}{N}}^{\frac{j}{N}} \left| f_y\left(\frac{j-1}{N}\right) - f_{\tilde{y}}\left(\frac{j-1}{N}\right) \right.$$
$$\left. + N\left(x - \frac{j-1}{N}\right)\left(y_j - \tilde{y}_j - f_y\left(\frac{j-1}{N}\right) + f_{\tilde{y}}\left(\frac{j-1}{N}\right)\right) \right| dx \tag{194}$$

$$= \int_{\frac{j-1}{N}}^{\frac{j}{N}} \left| N\left(x - \frac{j-1}{N}\right)(y_j - \tilde{y}_j) \right| dx \tag{195}$$

$$\geq \int_{\frac{j-1}{N}}^{\frac{j}{N}} \left| N\left(x - \frac{j-1}{N}\right)\frac{E}{M} \right| dx \tag{196}$$

$$= \frac{E}{2MN}. \tag{197}$$

where in (194) we used (190), (195) follows from $f_y(\frac{j-1}{N}) = f_{\tilde{y}}(\frac{j-1}{N})$, and in (196) we used $|y_j - \tilde{y}_j| \geq \frac{E}{M}$. Set $M = \lceil \frac{E}{4\varepsilon N} \rceil$. As $\varepsilon < \frac{E}{4N}$ by assumption, we have $\frac{E}{4\varepsilon N} > 1$ and hence $M = \lceil \frac{E}{4\varepsilon N} \rceil < \frac{E}{2\varepsilon N}$, where we used $\lceil x \rceil < 2x$, for $x > 1$. We therefore get $\frac{E}{2MN} > \varepsilon$, which, owing to (191)-(197), establishes that $\mathcal{F}_{N, \lceil \frac{E}{4\varepsilon N} \rceil}$ is an ε-packing of $\Sigma(X_N, E)$ with respect to the $L^1([0,1])$-norm. The proof is concluded by noting that

$$M(\varepsilon, \Sigma(X_N, E), L^1([0,1])) \geq \left| \mathcal{F}_{N, \lceil \frac{E}{4\varepsilon N} \rceil} \right| = \left| \left(\left\{ \frac{\ell}{M} E \right\}_{\ell=0}^{\lceil \frac{E}{4\varepsilon N} \rceil} \right)^N \right| \geq \left(\left\lceil \frac{E}{4\varepsilon N} \right\rceil \right)^N. \tag{198}$$

B.4 Proof of Lemma 7.3

We first need a concept closely related to fat-shattering dimension.

Definition B.2. *[9, Definition 2] Let $\mathcal{X}$ be a set and $\mathcal{F}$ a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The pseudodimension of $\mathcal{F}$, written as $Pdim(\mathcal{F})$, is the largest integer m for which there exists $(x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for every $(b_1, \ldots, b_m) \in \{0, 1\}^m$, there is an $f \in \mathcal{F}$ so that, for all $i \in \{1, \ldots, m\}$,*

$$f(x_i) \begin{cases} > y_i, & \text{if } b_i = 1, \\ \leq y_i, & \text{if } b_i = 0. \end{cases} \tag{199}$$

As condition (199) defining pseudodimension is weaker than condition (157) defining fat-shattering dimension, we have, for all function classes $\mathcal{F}$, that [16, Theorem 11.13 (i)]

$$\text{fat}(\mathcal{F}, \gamma) \leq \text{Pdim}(\mathcal{F}), \quad \text{for } \gamma \in \mathbb{R}_+. \tag{200}$$

We are now ready to show how Lemma 7.3 can be proved by applying results from [9]. First, note that [9] applies to families of network realizations whose associated configurations have a fixed architecture, whereas $\mathcal{N}(d, W, L)$, the object of interest here, consists of network configurations with different architectures. To resolve this discrepancy, we employ an idea used in the proof of [18, Lemma A.2]. Specifically, we consider the set $\mathcal{N}^*(d, 2W, L) = \{(A_\ell, b_\ell)_{\ell=1}^L : A_1 \in \mathbb{R}^{2W \times d}, b_1 \in \mathbb{R}^{2W}, A_L \in \mathbb{R}^{1 \times 2W}, b_L \in \mathbb{R}, A_\ell \in \mathbb{R}^{2W \times 2W}, b_\ell \in \mathbb{R}^{2W}, \text{ for } \ell \in \{2, \ldots, L-1\}\}$ consisting of all network configurations with the (fixed) architecture

$$(N_\ell)_{\ell=0}^L = (d, \underbrace{2W, \ldots,}_{(L-1) \text{ times}} 1). \tag{201}$$

The associated family of network realizations is $\mathcal{R}^*(d, 2W, L) = \{R(\Phi) : \Phi \in \mathcal{N}^*(d, 2W, L)\}$. It now follows from the proof of [18, Lemma A.2] that $\mathcal{R}(d, W, L) \subseteq \mathcal{R}^*(d, 2W, L)$. Next, we note that the network configurations in $\mathcal{N}^*(d, 2W, L)$ have $n(d, 2W, L) := 2dW + 4W + 1 + (L-2)((2W)^2 + 2W)$ weights. As $\mathcal{R}^*(d, 2W, L)$ consists of realizations of network configurations with fixed architecture, namely (201), we can apply the results in [9]. Specifically, we obtain

$$\text{Pdim}(\mathcal{R}^*(d, 2W, L)) \leq Cn(d, 2W, L)L \log(n(d, 2W, L)) \tag{202}$$
$$\leq C(13W^2L)L \log(13W^2L) \tag{203}$$
$$\leq 65\, CW^2L^2(\log(WL)), \tag{204}$$

where $C \in \mathbb{R}_+$ is an absolute constant, in (202) we used [9, Eq. (2)] combined with the discussion at the end of the paragraph immediately after [9, Definition 2], (203) follows from $n(d, 2W, L) \leq 2W^2 + 4W + 1 + (L-2)((2W)^2 + 2W) \leq 13W^2L$ by the standing assumption $W \geq d$, and (204) is thanks to $\log(13W^2L) \leq \log((WL)^5) = 5(\log(WL))$ as $L \geq 2$. Since $\mathcal{R}(d, W, L) \subseteq \mathcal{R}^*(d, 2W, L)$, as noted above, we get

$$\text{Pdim}(\mathcal{R}(d, W, L)) \leq \text{Pdim}(\mathcal{R}^*(d, 2W, L)) \leq 65\, CW^2L^2(\log(WL)). \tag{205}$$

To upper-bound the fat-shattering dimension of $\mathcal{T}_1 \circ \mathcal{R}(d, W, L)$, we first note that $\mathcal{T}_1(x) = -1 + \rho(x+1) - \rho(x-1)$, for $x \in \mathbb{R}$, and hence $\mathcal{T}_1 \in \mathcal{R}(1, 2, 2)$, which upon application of [18, Lemma H.3], yields

$$\mathcal{T}_1 \circ \mathcal{R}(d, W, L) \subseteq \mathcal{R}(d, \max\{W, 2\}, L+2) = \mathcal{R}(d, W, L+2). \tag{206}$$

It then follows, for all $\gamma \in \mathbb{R}_+$, that

$$\text{fat}(\mathcal{T}_1 \circ \mathcal{R}(d, W, L), \gamma) \leq \text{Pdim}(\mathcal{T}_1 \circ \mathcal{R}(d, W, L)) \tag{207}$$
$$\leq \text{Pdim}(\mathcal{R}(d, W, L+2)) \tag{208}$$
$$\leq 65\, CW^2(L+2)^2(\log(W(L+2))) \tag{209}$$
$$< 520\, CW^2L^2(\log(WL)), \tag{210}$$

where (207) is (200) with $\mathcal{F} = \mathcal{T}_1 \circ \mathcal{R}(d, W, L)$, (208) follows from (206), in (209) we used (205) with L replaced by $L+2$, and (210) follows from $(L+2)^2 \leq (2L)^2$ and $\log(W(L+2)) \leq \log(WL^2) < 2\log(WL)$, recalling that $L \geq 2$. The proof is concluded upon taking $C_h = 520\, C$.

C Proof of Theorem 4.1

The proof will be effected by establishing a slightly stronger result; this is done to better illustrate the roles of the assumptions in Theorem 4.1. Specifically, we shall replace the assumption (78) by

$$\inf_{f \in \mathcal{F}_n} \|g - f\|_{L^2(P)} \leq A, \quad A \in \mathbb{R}_+, \tag{211}$$

and the assumption (79) by

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(x_i) - y_i)^2 \leq \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \right) + \kappa, \quad \text{a.s.}, \quad \kappa \in \mathbb{R}_+. \tag{212}$$

With this new set of assumptions, we have the following theorem.

Theorem C.1. *Let $\mathbb{X} \subseteq \mathbb{R}^d$ and consider the regression function $g : \mathbb{X} \mapsto \mathbb{R}$. Let $n \in \mathbb{N}$ and $\sigma \in \mathbb{R}_+$. Let P be a distribution on $\mathbb{X}$, with the associated samples $(x_i, y_i)_{i=1}^n = (x_i, g(x_i) + \sigma \xi_i)_{i=1}^n$, where $(x_i)_{i=1}^n$ are i.i.d. random variables of distribution P, $(\xi_i)_{i=1}^n$ are i.i.d. standard Gaussian random variables, and $(x_i)_{i=1}^n$ and $(\xi_i)_{i=1}^n$ are statistically independent.*

Let $A, \kappa \in \mathbb{R}_+$, and consider a class of functions $\mathcal{F}_n \subseteq L^\infty(\mathbb{X})$ such that

$$\inf_{f \in \mathcal{F}_n} \|g - f\|_{L^2(P)} \leq A, \tag{213}$$

and an $\mathcal{F}_n$-valued random variable $\hat{f}_n$ satisfying

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(x_i) - y_i)^2 \leq \inf_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \right) + \kappa, \quad \text{a.s.} \tag{214}$$

For all $\delta \in (0, 1/2)$, it holds that

$$E(\|\hat{f}_n - g\|^2_{L^2(P)})$$
$$\leq 16(A^2 + \kappa) + 64(\sigma + \delta)\delta + 800(\sigma + \sigma^2 + (R(g, \mathcal{F}_n))^2) \frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n},$$

where $R(g, \mathcal{F}_n) := \max\{\|g\|_{L^\infty(\mathbb{X})}, \sup_{f \in \mathcal{F}_n} \|f\|_{L^\infty(\mathbb{X})}\}$.

Taking $A = \varepsilon$, $\kappa = \varepsilon^2$, and $\delta = \varepsilon^2$, with $\varepsilon \in (0, 1/2)$, in Theorem C.1 implies Theorem 4.1 with $C = 800$. To prepare for the proof of Theorem C.1, we state two auxiliary technical lemmata. The first one provides an upper bound on the expected empirical risk.

Lemma C.2. *Let $\mathbb{X}$, P, g, n, σ, $(x_i, y_i)_{i=1}^n$, A, κ, $\mathcal{F}_n$, and $\hat{f}_n$ be defined as in Theorem C.1 and assume that (213) and (214) hold. For all $\delta \in (0, 1/2)$, we have*

$$E\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(x_i) - g(x_i))^2 \right)$$
$$\leq 2(A^2 + \kappa) + 8\sigma\delta + 100(\sigma + \sigma^2) \frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}.$$

Proof. See Appendix C.1. □

The second lemma relates the expected empirical risk to the expected prediction error.

Lemma C.3. *Let $\mathbb{X}$, P, g, n, σ, $(x_i, y_i)_{i=1}^n$, A, κ, $\mathcal{F}_n$, and $\hat{f}_n$ be defined as in Theorem C.1 and assume that (213) and (214) hold. For all $\delta \in (0, 1/2)$, we have*

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2)$$
$$\leq 8E\left(\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2\right) + 64\left((R(g, \mathcal{F}_n))^2 \frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})))}{n} + \delta^2\right),$$

where $R(g, \mathcal{F}_n) := \max\{\|g\|_{L^\infty(\mathbb{X})}, \sup_{f \in \mathcal{F}_n} \|f\|_{L^\infty(\mathbb{X})}\}$.

Proof. See Appendix C.2. $\qquad\square$

Putting Lemmata C.2 and C.3 together, we can now finalize the proof of Theorem C.1 as follows:

$$E(\|\hat{f}_n - g\|_{L^2(P)}) \tag{215}$$

$$\leq 8E\left(\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2\right) + 64\left((R(g, \mathcal{F}_n))^2 \frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})))}{n} + \delta^2\right) \tag{216}$$

$$\leq 16(A^2 + \kappa) + 64\sigma\delta + 800(\sigma + \sigma^2)\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}$$
$$+ 64(R(g, \mathcal{F}_n))^2 \frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})))}{n} + 64\delta^2 \tag{217}$$

$$\leq 16(A^2 + \kappa) + 64(\sigma + \delta)\delta + 800(\sigma + \sigma^2 + (R(g, \mathcal{F}_n))^2)\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}. \tag{218}$$

C.1 Proof of Lemma C.2

Arbitrarily fix $f \in \mathcal{F}_n$. By assumption (214), we have

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - y_i)^2 \leq \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 + \kappa, \quad \text{a.s.} \tag{219}$$

Substituting $y_i = g(x_i) + \sigma\xi_i$, $i = 1, \ldots, n$, into (219), yields

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i) - \sigma\xi_i)^2 \leq \frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i) - \sigma\xi_i)^2 + \kappa, \quad \text{a.s.} \tag{220}$$

which results in

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2 \leq \frac{1}{n}\sum_{i=1}^n \left((f(x_i) - g(x_i))^2 + \frac{2\sigma}{n}\xi_i(\hat{f}_n(x_i) - f(x_i))\right) + \kappa, \quad \text{a.s.} \tag{221}$$

Taking expectations in (221) yields

$$E\left(\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2\right) \tag{222}$$

$$\leq E\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - g(x_i))^2\right) + \frac{2\sigma}{n}E\left(\sum_{i=1}^n \xi_i(\hat{f}_n(x_i) - f(x_i))\right) + \kappa \tag{223}$$

$$= \|f - g\|_{L^2(P)}^2 + \frac{2\sigma}{n}E\left(\sum_{i=1}^n \xi_i(\hat{f}_n(x_i) - f(x_i))\right) + \kappa \tag{224}$$

$$= \|f - g\|_{L^2(P)}^2 + \frac{2\sigma}{n}E\left(\sum_{i=1}^n \xi_i(\hat{f}_n(x_i) - g(x_i))\right) + \kappa \tag{225}$$

where in (224) we used that, for $i = 1, \ldots, n$, x_i is a random variable of distribution P and hence $E((f(x_i) - g(x_i))^2) = \|f - g\|_{L^2(P)}^2$, and (225) follows from adding $\frac{2\sigma}{n} E(\sum_{i=1}^{n} \xi_i(f(x_i) - g(x_i))) = 0$ to (224). As the choice of $f \in \mathcal{F}_n$ was arbitrary, (222)-(225) holds for all $f \in \mathcal{F}_n$, and therefore

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(x_i) - g(x_i))^2\right) \leq \inf_{f \in \mathcal{F}_n} \|f - g\|_{L^2(P)}^2 + \frac{2\sigma}{n} E\left(\sum_{i=1}^{n} \xi_i(\hat{f}_n(x_i) - g(x_i))\right) + \kappa. \quad (226)$$

With $\inf_{f \in \mathcal{F}_n} \|g - f\|_{L^2(P)} \leq A$, owing to assumption (213), and setting

$$\Delta = (\hat{f}_n(x_i) - g(x_i))_{i=1}^{n}, \quad (227)$$
$$\xi = (\xi_i)_{i=1}^{n}, \quad (228)$$

it follows from (226) that

$$E\left(\frac{\|\Delta\|_2^2}{n}\right) \leq A^2 + \kappa + \frac{2\sigma}{n} E(\langle \xi, \Delta \rangle). \quad (229)$$

We next note that the quantity $E(\langle \xi, \Delta \rangle)$ can be upper-bounded by the expected supremum of a Gaussian process according to

$$E(\langle \xi, \Delta \rangle) = E(\langle \xi, (\hat{f}_n(x_i) - g(x_i))_{i=1}^{n} \rangle) \leq E\left(\sup_{f \in \mathcal{F}_n} \langle \xi, (f(x_i) - g(x_i))_{i=1}^{n} \rangle\right). \quad (230)$$

The right-hand-side of (230) can be further upper-bounded either in terms of the covering number of $\mathcal{F}_n$, through one-step discretization, or by using the more advanced Dudley entropy integral bound, see e.g. [6, Section 5.3]. The one-step discretization approach turns out to suffice for the purposes of this proof. Specifically, let $\{f_j\}_{j=1}^{N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}$ be a δ-covering of $\mathcal{F}_n$, with respect to the $L^\infty(\mathbb{X})$-norm, and define the associated set of random vectors $\{\Delta_j\}_{j=1}^{N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}$ according to

$$\Delta_j = (f_j(x_i) - g(x_i))_{i=1}^{n}.$$

Further, define the random vector Δ^p as the element in $\{\Delta_j\}_{j=1}^{N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}$ that is closest to Δ, in the sense of

$$\Delta^p = \operatorname*{argmin}_{\Delta_j : j=1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))} \|\Delta - \Delta_j\|_2, \quad (231)$$

and note that, a.s.,

$$\|\Delta^p - \Delta\|_2 = \min_{j=1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))} \|\Delta - \Delta_j\|_2 \quad (232)$$

$$= \min_{j=1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))} \left(\sum_{i=1}^{n} (\hat{f}_n(x_i) - f_j(x_i))^2\right)^{1/2} \quad (233)$$

$$\leq \min_{j=1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))} n^{1/2} \|\hat{f}_n - f_j\|_{L^\infty(\mathbb{X})} \quad (234)$$

$$\leq n^{1/2} \delta, \quad (235)$$

where in (235) we used the fact that $\{f_j\}_{j=1}^{N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}$ is a δ-covering of $\mathcal{F}_n$ with respect to the $L^\infty(\mathbb{X})$-norm. We now get

$$E(\langle \xi, \Delta \rangle) = E(\langle \xi, \Delta - \Delta^p \rangle) + E(\langle \xi, \Delta^p \rangle) \quad (236)$$

$$= E(\langle \xi, \Delta - \Delta^p \rangle) + E(\|\Delta^p\|_2 \langle \xi, \|\Delta^p\|_2^{-1} \Delta^p \rangle) \quad (237)$$

$$\leq E(\langle \xi, \Delta - \Delta^p \rangle) + \sqrt{E(\|\Delta^p\|_2^2)} \sqrt{E(\langle \xi, \|\Delta^p\|_2^{-1} \Delta^p \rangle^2)}, \quad (238)$$

where (238) is thanks to the Cauchy-Schwarz inequality. We shall next upper-bound the terms $E(\langle \xi, \Delta - \Delta^p \rangle)$, $E(\|\Delta^p\|_2^2)$, and $E(\langle \xi, \|\Delta^p\|_2^{-1}\Delta^p \rangle^2)$, individually. First, note that

$$E(\langle \xi, \Delta - \Delta^p \rangle) \leq \sqrt{E(\langle \xi, \xi \rangle)}\sqrt{E(\langle \Delta - \Delta^p, \Delta - \Delta^p \rangle)} \tag{239}$$

$$\leq n\delta, \tag{240}$$

where (240) follows from $E(\langle \xi, \xi \rangle) = E(\sum_{i=1}^n \xi_i^2) = n$ and $E(\Delta - \Delta^p, \Delta - \Delta^p) = E(\|\Delta - \Delta^p\|_2^2) \leq n\delta^2$ by (232)-(235). We proceed to upper-bound $E(\|\Delta^p\|_2^2)$ as follows

$$E(\|\Delta^p\|_2^2) \leq E((\|\Delta\|_2 + \|\Delta^p - \Delta\|_2)^2) \tag{241}$$

$$\leq E((\|\Delta\|_2 + n^{1/2}\delta)^2) \tag{242}$$

$$\leq E(2\|\Delta\|_2^2 + 2(n^{1/2}\delta)^2) \tag{243}$$

$$\leq 2E(\|\Delta\|_2^2) + 4\sqrt{E(\|\Delta\|_2^2)}(n^{1/2}\delta) + 2(n^{1/2}\delta)^2 \tag{244}$$

$$= 2\left(\sqrt{E(\|\Delta\|_2^2)} + n^{1/2}\delta\right)^2, \tag{245}$$

where in (242) we again used (232)-(235) and (243) follows from $(a + b)^2 \leq 2a^2 + 2b^2$, with $a = \|\Delta\|_2$ and $b = n^{1/2}\delta$. It remains to upper-bound the term $E(\langle \xi, \|\Delta^p\|_2^{-1}\Delta^p \rangle^2)$. To this end, we first note that

$$E(\langle \xi, \|\Delta^p\|_2^{-1}\Delta^p \rangle^2) \leq E\left(\max_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2\right). \tag{246}$$

The right-hand-side of (246) can now be further upper-bounded through the moment generating function method. Specifically, let $t \in (0, 1/2)$ be a parameter to be determined later. We have

$$E\left(\max_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2\right) \tag{247}$$

$$= \frac{1}{t}\ln\left(\exp\left(E\left(t \max_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2\right)\right)\right) \tag{248}$$

$$\leq \frac{1}{t}\ln\left(E\left(\exp\left(t \max_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2\right)\right)\right) \tag{249}$$

$$= \frac{1}{t}\ln\left(E\left(\max_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \exp(t\langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2)\right)\right) \tag{250}$$

$$\leq \frac{1}{t}\ln\left(E\left(\sum_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \exp(t\langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2)\right)\right) \tag{251}$$

$$= \frac{1}{t}\ln\left(\sum_{j=1,\dots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} E(\exp(t\langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2))\right), \tag{252}$$

where in (249) we applied the Jensen inequality. For $j = 1, \dots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))$, conditioned on $(x_i)_{i=1}^n$, $\|\Delta_j\|_2^{-1}\Delta_j = \frac{(f_j(x_i)-g(x_i))_{i=1}^n}{\|(f_j(x_i)-g(x_i))_{i=1}^n\|_2}$ is a deterministic vector of $\|\cdot\|_2$-norm equal to 1 and hence $\langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2$ is a χ^2 random variable with 1 degree of freedom. We therefore get

$$E(\exp(t\langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2)|(x_i)_{i=1}^n) = (1 - 2t)^{-1/2}, \quad \text{a.s.} \tag{253}$$

By the law of total expectation, it finally follows that

$$E(\exp(t\langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2)) = (1 - 2t)^{-1/2},$$

for $j = 1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))$, which together with (247)-(252) yields

$$E\left(\max_{j=1,\ldots,N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))} \langle \xi, \|\Delta_j\|_2^{-1}\Delta_j \rangle^2\right) \leq \frac{1}{t}\ln(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))(1-2t)^{-1/2}) \tag{254}$$

$$= \frac{\ln N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}{t} - \frac{\ln(1-2t)}{2t}. \tag{255}$$

Putting (246) and (254)-(255) with $t = \frac{1}{10}$ together, we obtain

$$E(\langle \xi, \|\Delta^p\|_2^{-1}\Delta^p \rangle^2) \leq \frac{\ln N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}{1/10} - \frac{\ln(4/5)}{1/5} \leq 10\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 2. \tag{256}$$

Next, we have

$$E\left(\frac{\|\Delta\|_2^2}{n}\right) \tag{257}$$

$$\leq A^2 + \kappa + \frac{2\sigma}{n}E(\langle \xi, \Delta \rangle) \tag{258}$$

$$\leq A^2 + \kappa + \frac{2\sigma}{n}\left(E(\langle \xi, \Delta - \Delta^p \rangle) + \sqrt{E(\|\Delta^p\|_2^2)}\sqrt{E(\langle \xi, \|\Delta^p\|_2^{-1}\Delta^p \rangle^2)}\right) \tag{259}$$

$$\leq A^2 + \kappa + \frac{2\sigma}{n}\left(n\delta + \sqrt{2}\left(\sqrt{E(\|\Delta\|_2^2)} + n^{1/2}\delta\right)\sqrt{10\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 2}\right) \tag{260}$$

$$\leq A^2 + \kappa + \frac{2\sigma}{n}\left(n\delta + 5\left(\sqrt{E(\|\Delta\|_2^2)} + n^{1/2}\delta\right)\sqrt{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}\right) \tag{261}$$

$$= A^2 + \kappa + 2\sigma\delta + 10\sigma\sqrt{\frac{E(\|\Delta\|_2^2)}{n}}\sqrt{\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}}$$
$$+ 10\sigma\delta\sqrt{\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}}, \tag{262}$$

where (258) is (229), in (259) we used (236)-(238), and (260) follows from (239)-(240), (241)-(245), and (256). Regarding the last two terms in (262), we note that first

$$10\sigma\sqrt{\frac{E(\|\Delta\|_2^2)}{n}}\sqrt{\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}}$$
$$\leq \frac{E(\|\Delta\|_2^2)}{2n} + 50\sigma^2\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}, \tag{263}$$

where we used $ab \leq \frac{a^2+b^2}{2}$, with $a = \sqrt{\frac{E(\|\Delta\|_2^2)}{n}}$ and $b = 10\sigma\sqrt{\frac{\log(N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))+1}{n}}$, and second

$$10\sigma\delta\sqrt{\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}} \leq 2\sigma\delta^2 + \frac{25}{2}\sigma\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}, \tag{264}$$

which again is by $ab \leq \frac{a^2+b^2}{2}$, here with $a = 2\sigma^{1/2}\delta$ and $b = 5\sigma^{1/2}\sqrt{\frac{\log(N(\delta,\mathcal{F}_n,L^\infty(\mathbb{X}))+1}{n}}$. Substituting (263) and (264) into (257)-(262), rearranging terms, and using $\delta^2 \leq \delta$, yields

$$E\left(\frac{\|\Delta\|_2^2}{n}\right) \tag{265}$$

$$= 2A^2 + 2\kappa + 4\sigma\delta + 4\sigma\delta^2 + 100\sigma^2\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}$$
$$+ 25\sigma\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n} \tag{266}$$

$$\leq 2(A^2 + \kappa) + 8\sigma\delta + 100(\sigma + \sigma^2)\frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1}{n}. \tag{267}$$

The proof is concluded upon noting that $E(\frac{\|\Delta\|_2^2}{n}) = E(\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2)$.

C.2 Proof of Lemma C.3

We start with one-step discretization [6, Section 5.3] for $\mathcal{F}_n$. Let $\{f_j\}_{j=1}^{N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}$ be a δ-covering of $\mathcal{F}_n$ with respect to the $L^\infty(\mathbb{X})$-norm. Then, for each $f \in \mathcal{F}_n$, there exists an index $j(f) \in \{1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))\}$ such that

$$\|f - f_{j(f)}\|_{L^\infty(\mathbb{X})} \le \delta. \tag{268}$$

Next, we have, a.s.,

$$\|\hat{f}_n - g\|_{L^2(P)}^2 = \int (\hat{f}_n(x) - g(x))^2 dP(x) \tag{269}$$

$$\le \int \left(2(\hat{f}_n(x) - f_{j(\hat{f}_n)}(x))^2 + 2(f_{j(\hat{f}_n)}(x) - g(x))^2 \right) dP(x) \tag{270}$$

$$= 2\|\hat{f}_n - f_{j(\hat{f}_n)}\|_{L^2(P)}^2 + 2\|f_{j(\hat{f}_n)} - g\|_{L^2(P)}^2 \tag{271}$$

$$\le 2\delta^2 + 2\|f_{j(\hat{f}_n)} - g\|_{L^2(P)}^2, \tag{272}$$

where in (270) we used $(a+b)^2 \le 2a^2 + 2b^2$, and (272) follows from (268). Taking expectations in (269)-(272), yields

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) \le 2\delta^2 + 2E(\|f_{j(\hat{f}_n)} - g\|_{L^2(P)}^2). \tag{273}$$

Moreover, we have

$$E\left(\frac{1}{n} \sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2 \right) \tag{274}$$

$$\ge E\left(\frac{1}{n} \sum_{i=1}^n \left(\frac{1}{2}(f_{j(\hat{f}_n)}(x_i) - g(x_i))^2 - (f_{j(\hat{f}_n)}(x_i) - \hat{f}_n(x_i))^2 \right) \right) \tag{275}$$

$$\ge \frac{1}{2} E\left(\frac{1}{n} \sum_{i=1}^n (f_{j(\hat{f}_n)}(x_i) - g(x_i))^2 \right) - \delta^2, \tag{276}$$

where (275) follows from $a^2 \ge \frac{1}{2}(a+b)^2 - b^2$, with $a = \hat{f}_n(x_i) - g(x_i)$ and $b = f_{j(\hat{f}_n)}(x_i) - \hat{f}_n(x_i)$, for $i = 1, \ldots, n$, and in (276) we used (268). We next note that

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) - 8E\left(\frac{1}{n} \sum_{i=1}^n (\hat{f}_n(x_i) - g(x_i))^2 \right) \tag{277}$$

$$\le 2\delta^2 + 2E(\|f_{j(\hat{f}_n)} - g\|_{L^2(P)}^2) - 4E\left(\frac{1}{n} \sum_{i=1}^n (f_{j(\hat{f}_n)}(x_i) - g(x_i))^2 \right) + 8\delta^2 \tag{278}$$

$$= 2E\left(\|f_{j(\hat{f}_n)} - g\|_{L^2(P)}^2 - \frac{2}{n} \sum_{i=1}^n (f_{j(\hat{f}_n)}(x_i) - g(x_i))^2 \right) + 10\delta^2 \tag{279}$$

$$\le 2E\left(\sup_{j=1,\ldots,N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))} \left(\|f_j - g\|_{L^2(P)}^2 - \frac{2}{n} \sum_{i=1}^n (f_j(x_i) - g(x_i))^2 \right) \right) + 10\delta^2, \tag{280}$$

where (278) follows from (273) and (274)-(276). To simplify notation, in the following, we let

$$Z_{j,i} = \frac{(f_j(x_i) - g(x_i))^2}{4(R(g, \mathcal{F}_n))^2}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})). \tag{281}$$

Next, note that, for fixed j, $\{Z_{j,i}\}_{i=1}^{n}$ are i.i.d nonnegative random variables with mean

$$\mu_j = E(Z_{j,1}) = \int \frac{(f_j(x) - g(x))^2}{4(R(g, \mathcal{F}_n))^2} dP(x) = \frac{1}{4(R(g, \mathcal{F}_n))^2} \|f_j - g\|_{L^2(P)}^2. \tag{282}$$

Moreover, for $j = 1, \ldots, N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))$, $i = 1, \ldots, n$, we have

$$Z_{j,i} \in [0,1], \quad \text{a.s.} \tag{283}$$

as a consequence of $\|f_j - g\|_{L^\infty(\mathbb{X})} \le \|f_j\|_{L^\infty(\mathbb{X})} + \|g\|_{L^\infty(\mathbb{X})} \le 2R(g, \mathcal{F}_n)$. We can now rewrite (277)-(280) as

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) - 8E\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(x_i) - g(x_i))^2\right) \tag{284}$$

$$\le 8(R(g, \mathcal{F}_n))^2 E\left(\sup_{j=1,\ldots,N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}\left(\mu_j - \frac{2}{n}\sum_{i=1}^{n} Z_{j,i}\right)\right) + 10\,\delta^2 \tag{285}$$

and use the following lemma to upper-bound $E(\sup_{j=1,\ldots,N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}(\mu_j - \frac{2}{n}\sum_{i=1}^{n} Z_{j,i}))$.

Lemma C.4. *Let $U, V \in \mathbb{N}$. For $j = 1, \ldots, U$, let $\{Z_{j,i}\}_{i=1}^{V}$ be i.i.d nonnegative random variables of means μ_j and taking values in $[0,1]$ a.s. Then,*

$$E\left(\sup_{j=1,\ldots,U}\left(\mu_j - \frac{2}{V}\sum_{i=1}^{V} Z_{j,i}\right)\right) \le \frac{8\log(U)}{V}. \tag{286}$$

Proof. See Section C.3. $\qquad\square$

Application of Lemma C.4 with $U = N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))$, $V = n$, and the prerequisite satisfied thanks to (283), yields

$$E\left(\sup_{j=1,\ldots,N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X}))}\left(\mu_j - \frac{2}{n}\sum_{i=1}^{n} Z_{j,i}\right)\right) \le \frac{8\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})))}{n}, \tag{287}$$

which together with (284)-(285) finishes the proof of Lemma C.3 according to

$$E(\|\hat{f}_n - g\|_{L^2(P)}^2) - 8E\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(x_i) - g(x_i))^2\right)$$

$$\le 8(R(g, \mathcal{F}_n))^2 \frac{8\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})))}{n} + 10\delta^2$$

$$\le 64\left((R(g, \mathcal{F}_n))^2 \frac{\log(N(\delta, \mathcal{F}_n, L^\infty(\mathbb{X})))}{n} + \delta^2\right).$$

We finally note that the standard upper bound on sub-Gaussian maxima, as e.g. in [6, Exercise 2.12], would yield $E(\sup_{j=1,\ldots,U}(\mu_j - \frac{1}{V}\sum_{i=1}^{V} Z_{j,i})) \le C\sqrt{\frac{\log(U)}{V}}$, for some absolute constant C. Identifying $\log(U)/V$ with the right-most term in the prediction error upper bound (80) in Theorem 4.1, given by $(\log(N(\varepsilon^2, \mathcal{F}_n, L^\infty(\mathbb{X}))) + 1)/n$, as was essentially done above to finish the proof of Lemma C.3, we can see that as the prediction error upper bound goes to zero, we will be in the regime $\log(U)/V < 1$. In this case Lemma C.4 provides an upper bound that is stronger order-wise than the standard bound on sub-Gaussian maxima. This improvement is fundamental in our context as we will want the prediction error to approach zero at a certain rate.

C.3 Proof of Lemma C.4

Let

$$Q := \sup_{j=1,\ldots,U} \left(\mu_j - \frac{2}{V} \sum_{i=1}^{V} Z_{j,i} \right) \tag{288}$$

$$= 4 \sup_{j=1,\ldots,U} \left(\frac{1}{V} \sum_{i=1}^{V} \frac{1}{2}(\mu_j - Z_{j,i}) - \frac{1}{4}\mu_j \right). \tag{289}$$

For $j \in \{1, \ldots, U\}$, let

$$T_{j,i} = \frac{1}{2}(\mu_j - Z_{j,i}), \quad i = 1, \ldots, V, \tag{290}$$

and note that, for fixed j, $\{T_{j,i}\}_{i=1}^{V}$ are i.i.d. random variables with

$$E(T_{j,i}) = E\left(\frac{1}{2}\left(\mu_j - Z_{j,i} \right) \right) = 0, \tag{291}$$

$$|T_{j,i}| \leq \frac{1}{2}(|\mu_j| + |Z_{j,i}|) \leq 1, \quad \text{a.s.} \tag{292}$$

and variance

$$\sigma_j^2 = E\left(\left(\frac{1}{2}(\mu_j - Z_{j,i}) \right)^2 \right) \tag{293}$$

$$= \frac{1}{4}E((Z_{j,i}^2) - \frac{1}{4}\mu_j^2 \tag{294}$$

$$\leq \frac{1}{4}E(Z_{j,i}) \tag{295}$$

$$\leq \frac{1}{4}\mu_j, \tag{296}$$

where (295) is by $Z_{j,i} \in [0,1]$ a.s. With

$$S := \sup_{j=1,\ldots,U} \left(\frac{1}{V} \sum_{i=1}^{V} T_{j,i} - \sigma_j^2 \right),$$

it follows from (293)-(296) that

$$Q \leq 4S, \quad \text{a.s.} \tag{297}$$

We next upper-bound $E(S)$ by upper-bounding its moment generating function $\exp(E(tS)), t \in (0, V)$, and first note that

$$E(S) = \frac{1}{t} \ln(\exp(E(tS))) \leq \frac{1}{t} \ln(E(\exp(tS))), \tag{298}$$

where we used the Jensen inequality. Next, we have

$$E(\exp(tS)) = E\left(\exp\left(t \sup_{j=1,\ldots,U}\left(\frac{1}{V}\sum_{i=1}^{V}T_{j,i} - \sigma_j^2\right)\right)\right) \tag{299}$$

$$= E\left(\sup_{j=1,\ldots,U}\exp\left(t\left(\frac{1}{V}\sum_{i=1}^{V}T_{j,i} - \sigma_j^2\right)\right)\right) \tag{300}$$

$$\leq E\left(\sum_{j=1}^{U}\exp\left(t\left(\frac{1}{V}\sum_{i=1}^{V}T_{j,i} - \sigma_j^2\right)\right)\right) \tag{301}$$

$$= \sum_{j=1}^{U}E\left(\exp\left(t\left(\frac{1}{V}\sum_{i=1}^{V}T_{j,i} - \sigma_j^2\right)\right)\right) \tag{302}$$

$$= \sum_{j=1}^{U}E\left(\exp(-t\sigma_j^2)\prod_{i=1}^{V}\exp\left(\frac{t}{V}T_{j,i}\right)\right) \tag{303}$$

$$= \sum_{j=1}^{U}\exp(-t\sigma_j^2)\prod_{i=1}^{V}E\left(\exp\left(\frac{t}{V}T_{j,i}\right)\right). \tag{304}$$

To upper-bound $E(\exp(\frac{t}{V}T_{j,i}))$, for $j = 1,\ldots,U$, $i = 1,\ldots,V$, we note that, for $\lambda \in [0,1)$,

$$E\left(\exp\left(\lambda T_{j,i}\right)\right) = E\left(\sum_{k=0}^{\infty}\frac{(\lambda T_{j,i})^k}{k!}\right) \tag{305}$$

$$= E\left(1 + \lambda T_{j,i} + \sum_{k=2}^{\infty}\frac{(\lambda T_{j,i})^k}{k!}\right) \tag{306}$$

$$\leq 1 + E\left(\sum_{k=2}^{\infty}\frac{\lambda^k (T_{j,i})^2}{k!}\right) \tag{307}$$

$$= 1 + E\left(T_{j,i}^2\right)\sum_{k=2}^{\infty}\frac{\lambda^k}{k!} \tag{308}$$

$$\leq 1 + \sigma_j^2\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{2} \tag{309}$$

$$\leq 1 + \frac{\sigma_j^2\lambda^2}{2(1-\lambda)} \tag{310}$$

$$\leq \exp\left(\frac{\sigma_j^2\lambda^2}{2(1-\lambda)}\right), \tag{311}$$

where (307) follows from $|T_{j,i}| \leq 1$ a.s, in (310) we used $\sum_{k=0}^{\infty}\lambda^k = \frac{1}{1-\lambda}$, for $\lambda \in [0,1)$, and (311) is by $1 + x \leq \exp(x)$, $x \in \mathbb{R}$. Using (305)-(311), with $\lambda = \frac{t}{V}$, in (304), we obtain, for $t \in (0,V)$,

$$E(\exp(tS)) \leq \sum_{j=1}^{U}\exp\left(-t\sigma_j^2 + V\left(\frac{\sigma_j^2(\frac{t}{V})^2}{2(1-\frac{t}{V})}\right)\right)$$

$$= \sum_{j=1}^{U}\exp\left(V\sigma_j^2\left(-\frac{t}{V} + \left(\frac{(\frac{t}{V})^2}{2(1-\frac{t}{V})}\right)\right)\right),$$

which, by setting $t = \frac{V}{2}$, yields

$$E\left(\exp\left(\frac{V}{2}S\right)\right) \le \sum_{j=1}^{U} \exp\left(-\frac{V\sigma_j^2}{4}\right) \le U. \tag{312}$$

Finally, using (312) in (298), with $t = \frac{V}{2}$, we obtain

$$E(S) \le \frac{2\ln(U)}{V} \le \frac{2\log(U)}{V}. \tag{313}$$

The proof is concluded upon noting that

$$E\left(\sup_{j=1,\dots,U}\left(\mu_j - \frac{2}{V}\sum_{i=1}^{V} Z_{j,i}\right)\right) = E(Q) \le 4E(S) \le \frac{8\log(U)}{V},$$

where $Q \le 4S$ a.s. is by (297).

D   Proof of Theorem 5.1

We prove the upper bound (137) and the lower bound (138) in Appendices D.1 and D.2, respectively. Before delving into the proofs, we note that it suffices to consider the case $s < L(W^2 + W)$ as otherwise the network would qualify as fully connected, i.e., all weights may be nonzero, and Theorem 2.1 applies. Further, for simplicity of exposition and consistency with Theorem 2.1, we decided to work with the quantity LW^2 throughout as opposed to $L(W^2+W)$, simply by using $LW^2 \le L(W^2 + W) \le 2LW^2$.

D.1   Proof of the Upper Bound (137)

The overall proof architecture is identical to that of the upper bound (3) in Theorem 2.1. Specifically, we construct an explicit ε-covering of $\mathcal{R}(d,W,L,B,s)$ with elements in $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s)$[7], where $b \in \mathbb{N}$ is a parameter suitably depending on ε and determined later.

The proof requires two preparatory technical elements, the first of which, namely Lemma D.1, is very similar to Lemma 2.3.

Lemma D.1. *Let $p \in [1,\infty]$, $d,W,L,s,b \in \mathbb{N}$, and $B \in \mathbb{R}_+$ with $B \ge 1$. Then, the set $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s)$ is an $(L(W+1)^L B^{L-1} 2^{-b})$-covering of $\mathcal{R}(d,W,L,B,s)$ with respect to the $L^p([0,1]^d)$-norm.*

Proof. Let $q_b : [-B,B] \mapsto [-B,B] \cap 2^{-b}\mathbb{Z}$ be defined as

$$q_b(x) = \begin{cases} 2^{-b}\lfloor 2^b x\rfloor, & \text{for } x \in [0,B], \\ 2^{-b}\lceil 2^b x\rceil, & \text{for } x \in [-B,0), \end{cases}$$

and note that $|x - q_b(x)| \le 2^{-b}$, for all $x \in [-B,B]$. Arbitrarily fix $f \in \mathcal{R}(d,W,L,B,s)$. By definition, there exists $\Phi = ((A_\ell,b_\ell))_{\ell=1}^{\widetilde{L}} \in \mathcal{N}(d,W,L,B,s)$, with $\widetilde{L} \le L$, such that $R(\Phi) = f$. We now quantize the weights of Φ according to[8]

$$Q_b(\Phi) = ((q_b(A_\ell),q_b(b_\ell)))_{\ell=1}^{\widetilde{L}} \in \mathcal{N}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s),$$

---

[7]We note that $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s) = \mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,B,s)$. We chose, however, to use the notation $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s)$ for consistency with later parts of the proof involving $\mathcal{R}_{\mathbb{A}}(d,W,L,\infty,s)$ with general $\mathbb{A} \subseteq \mathbb{R}$.

[8]Note that $\mathcal{N}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s) = \mathcal{N}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,B,s)$. We chose, however, to use the notation $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d,W,L,\infty,s)$ for consistency with later parts of the proof.

where q_b acts elementwise. Here, we used the fact that, owing to $q_b(0) = 0$, the connectivity of $Q_b(\Phi)$ is no greater than that of Φ. Next, note that

$$\|\Phi - Q_b(\Phi)\| = \max_{\ell=1,\ldots,\widetilde{L}} \max\{\|A_\ell - q_b(A_\ell)\|_\infty, \|b_\ell - q_b(b_\ell)\|_\infty\} \leq 2^{-b},$$

which together with Lemma 2.2 yields

$$\|R(\Phi) - R(Q_b(\Phi))\|_{L^\infty([0,1]^d)} \leq L(W+1)^L B^{L-1}\|\Phi - Q_b(\Phi)\| \leq L(W+1)^L B^{L-1} 2^{-b}. \quad (314)$$

As

$$\|R(\Phi) - R(Q_b(\Phi))\|_{L^p([0,1]^d)} \leq \sup_{x\in[0,1]^d} |R(\Phi)(x) - R(Q_b(\Phi))(x)| = \|R(\Phi) - R(Q_b(\Phi))\|_{L^\infty([0,1]^p)},$$

it follows from (314) that

$$\|R(\Phi) - R(Q_b(\Phi))\|_{L^p([0,1]^d)} \leq L(W+1)^L B^{L-1} 2^{-b}. \quad (315)$$

As $f \in \mathcal{R}(d, W, L, B, s)$ was chosen arbitrarily and $R(Q_b(\Phi)) \in \mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d, W, L, \infty, s) \subseteq \mathcal{R}(d, W, L, B, s)$, we can conclude that $\mathcal{R}_{[-B,B]\cap 2^{-b}\mathbb{Z}}(d, W, L, \infty, s)$ is an $(L(W+1)^L B^{L-1} 2^{-b})$-covering of $\mathcal{R}(d, W, L, B, s)$ with respect to the $L^p([0,1]^d)$-norm. $\qquad\square$

In the second preparatory step, we upper-bound the cardinality of the covering just identified, specifically we shall consider sets $\mathcal{R}_\mathbb{A}(d, W, L, \infty, s)$ with general $\mathbb{A}$ and then particularize the result to $\mathbb{A} = [-B, B] \cap 2^{-b}\mathbb{Z}$ below.

Lemma D.2. *Let $d, W, L, s \in \mathbb{N}$, and $\mathbb{A} \subseteq \mathbb{R}$, with $s \geq \max\{W, L\}$ and $|\mathbb{A}| \geq 2$. Then,*

$$\begin{aligned} \log(|\mathcal{R}_\mathbb{A}(d, W, L, \infty, s)|) &\leq \log(|\mathcal{N}_\mathbb{A}(d, W, L, \infty, s)|) \\ &\leq 5s(\log(L(W+1)) + \log(|\mathbb{A}|)). \end{aligned} \quad (316)$$

Proof. By definition,

$\mathcal{N}_\mathbb{A}(d, W, L, \infty, s)$
$= \{(A_\ell, b_\ell)_{\ell=1}^{\widetilde{L}} \in \mathcal{N}(d) : \mathcal{W}((A_\ell, b_\ell)_{\ell=1}^{\widetilde{L}}) \leq W, \widetilde{L} \leq L, \mathcal{M}((A_\ell, b_\ell)_{\ell=1}^{\widetilde{L}}) \leq s, \mathrm{coef}((A_\ell, b_\ell)_{\ell=1}^{\widetilde{L}}) \subseteq \mathbb{A}\}.$

There are at most $\sum_{\widetilde{L}=1}^{L} W^{\widetilde{L}} \leq LW^L$ different architectures $(N_0, \ldots, N_{\widetilde{L}})$[9], $\widetilde{L} \leq L$, for network configurations in $\mathcal{N}_\mathbb{A}(d, W, L, \infty, s)$. For a given architecture $(N_0, \ldots, N_{\widetilde{L}})$, the total number of weights, including zero and nonzero ones, satisfies $\sum_{\ell=1}^{\widetilde{L}}(N_\ell N_{\ell-1} + N_\ell) \leq L(W^2 + W)$ and there are at most

$$\sum_{i=0}^{s} \binom{L(W^2 + W)}{i} \leq \sum_{i=0}^{s} (L(W^2 + W))^i \leq (L(W^2 + W))^{s+1}$$

different ways to choose the positions of the nonzero weights in the network configuration. Finally, given an architecture and the positions of the nonzero weights, there are at most $|\mathbb{A}|^s$ different ways to choose these nonzero weights. It therefore follows that

$$|\mathcal{N}_\mathbb{A}(d, W, L, \infty, s)| \leq LW^L \cdot (L(W^2 + W))^{s+1} \cdot |\mathbb{A}|^s, \quad (317)$$

---

[9]Note that $N_0 = d$ is fixed.

43

and hence

$$\log(|\mathcal{N}_{\mathbb{A}}(d, W, L, \infty, s)|) \tag{318}$$
$$\leq \log(L) + L\log(W) + (s+1)\log(L(W^2 + W)) + s\log(|\mathbb{A}|) \tag{319}$$
$$\leq s\log(L(W+1)) + (s+1)\log(L(W^2 + W)) + s\log(|\mathbb{A}|) \tag{320}$$
$$\leq s\log(L(W+1)) + 4s\log(L(W+1)) + s\log(|\mathbb{A}|) \tag{321}$$
$$\leq 5s(\log(L(W+1)) + \log(|\mathbb{A}|)). \tag{322}$$

where (320) follows from $\log(L) + L\log(W) \leq s\log(L) + s\log(W) \leq s\log(L(W+1))$ and in (321) we used $(s+1) \leq 2s$ and $L(W^2 + W) \leq (L(W+1))^2$. The result then follows from (318)-(322) together with $|\mathcal{R}_{\mathbb{A}}(d, W, L, \infty, s)| \leq |\mathcal{N}_{\mathbb{A}}(d, W, L, \infty, s)|$. $\square$

We are now ready to prove the upper bound (137). Fix $\varepsilon \in (0, 1/2)$, let

$$b := \left\lceil \log\left(\frac{L(W+1)^L B^{L-1}}{\varepsilon}\right) \right\rceil, \tag{323}$$

and note that $L(W+1)^L B^{L-1} 2^{-b} \leq \varepsilon$. It follows from Lemma D.1 that $\mathcal{R}_{[-B,B] \cap 2^{-b}\mathbb{Z}}(d, W, L, \infty, s)$ is an ε-covering of $\mathcal{R}(d, W, L, B, s)$ with respect to the $L^p([0,1]^d)$-norm. By the minimality of the covering number, we have

$$N(\varepsilon, \mathcal{R}(d, W, L, B, s), L^p([0,1]^d)) \leq |\mathcal{R}_{[-B,B] \cap 2^{-b}\mathbb{Z}}(d, W, L, \infty, s)|. \tag{324}$$

It then follows that

$$\log(N(\varepsilon, \mathcal{R}(d, W, L, B, s), L^p([0,1]^d))) \tag{325}$$
$$\leq \log(|\mathcal{R}_{[-B,B] \cap 2^{-b}\mathbb{Z}}(d, W, L, \infty, s)|) \tag{326}$$
$$\leq 5s(\log(L(W+1)) + \log(|[-B,B] \cap 2^{-b}\mathbb{Z}|)) \tag{327}$$
$$\leq 5s\left(\log(L(W+1)) + 3\log\left(\frac{L(W+1)^L B^L}{\varepsilon}\right)\right) \tag{328}$$
$$\leq 5s\left(2\log\left(\frac{(W+1)^L B^L}{\varepsilon}\right) + 6\log\left(\frac{(W+1)^L B^L}{\varepsilon}\right)\right) \tag{329}$$
$$= 40s\log\left(\frac{(W+1)^L B^L}{\varepsilon}\right) \tag{330}$$
$$= 40\min\{s, 2W^2 L\}\log\left(\frac{(W+1)^L B^L}{\varepsilon}\right) \tag{331}$$
$$\leq 80\min\{s, W^2 L\}\log\left(\frac{(W+1)^L B^L}{\varepsilon}\right). \tag{332}$$

where (327) is by Lemma D.2, (328) follows from (13)-(18), and in (329) we used $L(W+1) \leq (W+1)^L (W+1) \leq \frac{(W+1)^{2L} B^{2L}}{\varepsilon^2}$ and $\frac{L(W+1)^L B^L}{\varepsilon} \leq \frac{(W+1)^{2L} B^{2L}}{\varepsilon^2}$. This concludes the proof upon taking $C = 80$.

D.2   Proof of the Lower Bound (138)

Set $D := 60^2 \cdot 6$, define the auxiliary variable

$$\overline{W} = \left\lfloor \sqrt{\frac{s}{6L}} \right\rfloor, \tag{333}$$

and note that $\overline{W} \leq W$ as a consequence of $\sqrt{\frac{s}{6L}} < \sqrt{\frac{2W^2L}{6L}} < W$. Further, we have $\overline{W} = \lfloor\sqrt{\frac{s}{6L}}\rfloor \geq \left\lfloor\sqrt{\frac{Dd^2L}{6L}}\right\rfloor = \left\lfloor\sqrt{\frac{60^2\cdot 6\cdot d^2 L}{6L}}\right\rfloor = 60d \geq \max\{60, d\}$. It now follows from $\overline{W} \leq W$ that

$$R(d, \overline{W}, L, B, s) \subseteq R(d, W, L, B, s). \tag{334}$$

As $2\overline{W}^2L = 2\lfloor\sqrt{\frac{s}{6L}}\rfloor^2 L \leq 2\cdot\frac{s}{6L}\cdot L < s$, we can conclude that

$$R(d, \overline{W}, L, B, s) = R(d, \overline{W}, L, B). \tag{335}$$

Combining (334) and (335), we obtain the inclusion

$$R(d, \overline{W}, L, B) \subseteq R(d, W, L, B, s),$$

which, thanks to (441) in Lemma F.1, yields

$$N(\varepsilon, \mathcal{R}(d, W, L, B, s), L^p([0,1]^d)) \geq N(2\varepsilon, \mathcal{R}(d, \overline{W}, L, B), L^p([0,1]^d)). \tag{336}$$

Application of (4) in Theorem 2.1 with W replaced by $\overline{W}$, ε replaced by 2ε, and the prerequisites satisfied owing to $\overline{W} \geq \max\{60, d\}$, $L \geq 60$, and $2\varepsilon \in (0, 1/2)$, yields a lower bound on the right-hand-side of (336) according to

$$\log(N(2\varepsilon, \mathcal{R}(d, \overline{W}, L, B), L^p([0,1]^d))) \geq c_1\overline{W}^2L\log\left(\frac{(\overline{W}+1)^L B^L}{2\varepsilon}\right), \tag{337}$$

with $c_1 \in \mathbb{R}_+$ an absolute constant. We can now further lower-bound the right-hand-side of (337) according to

$$c_1\overline{W}^2L\log\left(\frac{(\overline{W}+1)^L B^L}{2\varepsilon}\right) \tag{338}$$

$$\geq \frac{c_1}{24}s\log\left(\frac{(\overline{W}+1)^L B^L}{2\varepsilon}\right) \tag{339}$$

$$> \frac{c_1}{48}s\log\left(\frac{(\overline{W}+1)^L B^L}{\varepsilon}\right) \tag{340}$$

$$= \frac{c_1}{48}\min\{s, 2W^2L\}\log\left(\frac{(\overline{W}+1)^L B^L}{\varepsilon}\right) \tag{341}$$

$$\geq \frac{c_1}{48}\min\{s, W^2L\}\log\left(\frac{(\overline{W}+1)^L B^L}{\varepsilon}\right), \tag{342}$$

where (339) follows from $\overline{W}^2L = \lfloor\sqrt{\frac{s}{6L}}\rfloor^2 L \geq (\frac{1}{2}\sqrt{\frac{s}{6L}})^2 L \geq \frac{1}{24}s$, as $\lfloor x\rfloor \geq \frac{1}{2}x$, for $x \geq 1$, and (340) is by $\log(\frac{(\overline{W}+1)^L B^L}{2\varepsilon}) = \frac{1}{2}\log(\frac{(\overline{W}+1)^{2L} B^{2L}}{4\varepsilon^2}) > \frac{1}{2}\log(\frac{(\overline{W}+1)^L B^L}{\varepsilon})$, since $\varepsilon \in (0, \frac{1}{4})$. To replace $\overline{W}$ in (342) by $\widetilde{W} = \min\{\lceil\sqrt{\frac{s}{L}}\rceil, W\}$, we note that

$$(\overline{W}+1)^2 = (\overline{W}+1)\cdot\left(\left\lfloor\sqrt{\frac{s}{6L}}\right\rfloor + 1\right) \tag{343}$$

$$\geq 61\cdot\left\lceil\sqrt{\frac{s}{6L}}\right\rceil \tag{344}$$

$$> 10\sqrt{\frac{s}{L}} \tag{345}$$

$$> \left\lceil\sqrt{\frac{s}{L}}\right\rceil + 1 \tag{346}$$

$$\geq \widetilde{W} + 1, \tag{347}$$

where in (344) we used $\overline{W} \geq 60$. Using (343)-(347), we get

$$\log\left(\frac{(\overline{W}+1)^L B^L}{\varepsilon}\right) = \frac{1}{2}\log\left(\frac{(\overline{W}+1)^{2L} B^{2L}}{\varepsilon^2}\right) \geq \frac{1}{2}\log\left(\frac{(\widetilde{W}+1)^L B^L}{\varepsilon}\right). \tag{348}$$

Putting (336), (338)-(342), and (348) together, we obtain

$$\log(N(\varepsilon, \mathcal{R}(d, W, L, B, s), L^p([0,1]^d))) \geq \frac{c_1}{96} \min\{s, W^2 L\} \log\left(\frac{(\widetilde{W}+1)^L B^L}{\varepsilon}\right), \tag{349}$$

which concludes the proof upon setting $c = \frac{c_1}{96}$.

E Proof of the Lower Bound (140) in Theorem 6.1

The proof of the lower bound (140) relies on several technical ingredients, which we present first. We start with a result that allows to reduce the general case $a \in \mathbb{N}$ to $a = 1$.

Lemma E.1. *Let* $d, W, L, a, b \in \mathbb{N}$, *with* $W \geq 2$. *We have*

$$\mathcal{R}_b^a(d, W, L) = 2^{(a-1)L} \cdot \mathcal{R}_{a+b-1}^1(d, W, L) \tag{350}$$

$$= \{2^{(a-1)L} \cdot f : f \in \mathcal{R}_{a+b-1}^1(d, W, L)\}, \tag{351}$$

and, for all $\varepsilon \in \mathbb{R}_+$, $p \in [1, \infty]$,

$$N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d)) = N\left(\frac{\varepsilon}{2^{(a-1)L}}, \mathcal{R}_{a+b-1}^1(d, W, L), L^p([0,1]^d)\right). \tag{352}$$

Proof. We start by establishing (350). Arbitrarily fix $g \in \mathcal{R}_b^a(d, W, L)$. From Lemma F.2, with its prerequisite satisfied thanks to $\{-1, 0, 1\} \subseteq \mathbb{Q}_b^a$, we can infer the existence of a network configuration $\Phi = (A_\ell, b_\ell)_{\ell=1}^L$, with $\mathcal{W}(\Phi) \leq W$ and $\mathrm{coef}(\Phi) \in \mathbb{Q}_b^a$, such that $R(\Phi) = g$. Define $\widetilde{\Phi} = (\frac{1}{2^{a-1}} A_\ell, \frac{1}{2^{a-1}} b_\ell)_{\ell=1}^L$, and note that $\mathcal{W}(\widetilde{\Phi}) \leq W$, $\mathcal{L}(\widetilde{\Phi}) = L$, and $\mathrm{coef}(\widetilde{\Phi}) \subseteq \frac{1}{2^{a-1}}\mathbb{Q}_b^a = \mathbb{Q}_{b+a-1}^1$, and therefore $R(\widetilde{\Phi}) \in \mathcal{R}_{a+b-1}^1(d, W, L)$. We then have

$$g = R(\Phi) \tag{353}$$

$$= S(A_L, b_L) \circ \rho \circ \cdots \circ \rho \circ S(A_1, b_1) \tag{354}$$

$$= (2^{(a-1)})^L S\left(\frac{1}{2^{a-1}} A_L, \frac{1}{2^{a-1}} b_L\right) \circ \rho \circ \cdots \circ \rho \circ S\left(\frac{1}{2^{a-1}} A_1, \frac{1}{2^{a-1}} b_1\right) \tag{355}$$

$$= 2^{(a-1)L} \cdot R(\widetilde{\Phi}) \tag{356}$$

$$\in \{2^{(a-1)L} \cdot f : f \in \mathcal{R}_{a+b-1}^1(d, W, L)\}, \tag{357}$$

where in (355) we used the positive homogeneity of the ReLU function, namely, $\rho(kx) = k\rho(x)$ for all $x \in \mathbb{R}$ and $k \in \mathbb{R}_+$. As the choice of $g \in \mathcal{R}_b^a(d, W, L)$ was arbitrary, we have established that

$$\mathcal{R}_b^a(d, W, L) \subseteq 2^{(a-1)L} \cdot \mathcal{R}_{a+b-1}^1(d, W, L). \tag{358}$$

Upon noting that the reverse inclusion can be proved similarly, (350) follows.

We proceed to prove (352). Fix $\varepsilon \in \mathbb{R}_+$ and $p \in [1, \infty]$. It follows from (350) that, for every ε-covering $\mathcal{C}$ of $\mathcal{R}_b^a(d, W, L)$ with respect to the $L^p([0,1]^d)$-norm, $\frac{1}{2^{(a-1)L}} \cdot \mathcal{C}$ is an $\frac{\varepsilon}{2^{(a-1)L}}$-covering of $R_{a+b-1}^1(d, W, L)$, which allows us to conclude that

$$N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d)) \geq N\left(\frac{\varepsilon}{2^{(a-1)L}}, \mathcal{R}_{a+b-1}^1(d, W, L), L^p([0,1]^d)\right). \tag{359}$$

Moreover, based on (350), we can also conclude that, for every $\frac{\varepsilon}{2^{(a-1)L}}$-covering $\mathcal{C}$ of $\mathcal{R}^1_{a+b-1}(d,W,L)$ with respect to the $L^p([0,1]^d)$-norm, $2^{(a-1)L} \cdot \mathcal{C}$ is an ε-covering of $\mathcal{R}^a_b(d,W,L)$, which leads to

$$N\left(\frac{\varepsilon}{2^{(a-1)L}}, \mathcal{R}^1_{a+b-1}(d,W,L), L^p([0,1]^d)\right) \geq N(\varepsilon, \mathcal{R}^a_b(d,W,L), L^p([0,1]^d)). \qquad (360)$$

Combining (359) and (360), yields (352). □

We note that the constraint $W \geq 2$ in Lemma E.1 is not restrictive for the purposes of the proof of the lower bound in Theorem 6.1 as below we will take the absolute constant D in Theorem 6.1 to be greater than 2.

Next, we derive lower bounds on the covering number, separately, for large and moderate values of b. We start with the case of large b.

Lemma E.2. *Let $p \in [1,\infty]$ and $d,W,L,b \in \mathbb{N}$, with $W,L \geq 60$. Assume that $b > \log(L) + L\log(W+1) + 3$. Then, there exist absolute constants $c_1, c_2, c_3 \in \mathbb{R}_+$ such that the following statements hold:*

- *For all $\varepsilon \in (0, L(W+1)^L 2^{-b}]$,*

$$\log(N(\varepsilon, \mathcal{R}^1_b(d,W,L), L^p([0,1]^d))) \geq c_1 W^2 L b. \qquad (361)$$

- *If $L(W+1)^L 2^{-b} < \frac{1}{32}$, then, for all $\varepsilon \in (L(W+1)^L 2^{-b}, \frac{1}{32})$,*

$$\log(N(\varepsilon, \mathcal{R}^1_b(d,W,L), L^p([0,1]^d))) \geq c_2 W^2 L \log\left(\frac{(W+1)^L}{\varepsilon}\right). \qquad (362)$$

- *For all $\varepsilon \in (0, \frac{1}{32})$,*

$$\log(N(\varepsilon, \mathcal{R}^1_b(d,W,L), L^p([0,1]^d))) \geq c_3 W^2 L \cdot \min\left\{b, \log\left(\frac{(W+1)^L}{\varepsilon}\right)\right\}. \qquad (363)$$

Proof. We start with the proof of (361). It follows from Lemma 2.3 with $B = 1$ that

$$\mathcal{A}(\mathcal{R}(d,W,L,1), \mathcal{R}_{[-1,1]\cap 2^{-b}\mathbb{Z}}(d,W,L), \|\cdot\|_{L^p([0,1]^d)}) \leq L(W+1)^L 2^{-b}. \qquad (364)$$

As $([-1,1] \cap 2^{-b}\mathbb{Z}) \subseteq ((-4,4) \cap 2^{-b}\mathbb{Z}) = \mathbb{Q}^1_b$, we have $\mathcal{R}_{[-1,1]\cap 2^{-b}\mathbb{Z}}(d,W,L) \subseteq \mathcal{R}^1_b(d,W,L)$ which together with (364) yields

$$\begin{aligned} & \mathcal{A}(\mathcal{R}(d,W,L,1), \mathcal{R}^1_b(d,W,L), \|\cdot\|_{L^p([0,1]^d)}) \\ & \leq \mathcal{A}(\mathcal{R}(d,W,L,1), \mathcal{R}_{[-1,1]\cap 2^{-b}\mathbb{Z}}(d,W,L), \|\cdot\|_{L^p([0,1]^d)}) \\ & \leq L(W+1)^L 2^{-b}. \end{aligned} \qquad (365)$$

Application of Proposition 3.1 with $\delta = \|\cdot\|_{L^p([0,1]^d)}$, $\mathcal{G} = \mathcal{R}(d,W,L,1)$, $\mathcal{F} = \mathcal{R}^1_b(d,W,L)$, $\varepsilon = L(W+1)^L 2^{-b}$, and the prerequisite (39) satisfied thanks to (365), yields

$$\begin{aligned} & N(L(W+1)^L 2^{-b}, \mathcal{R}^1_b(d,W,L), L^p([0,1]^d)) \\ & \geq N(4L(W+1)^L 2^{-b}, \mathcal{R}(d,W,L,1), L^p([0,1]^d)). \end{aligned} \qquad (366)$$

The right-hand-side of (366) can be lower-bounded by application of (4) in Theorem 2.1 with $B = 1$, upon noting that the prerequisites are satisfied thanks to $W,L \geq 60$ by assumption, and $4L(W+1)^L 2^{-b} < 4L(W+1)^L 2^{-(\log(L)+L\log(W+1)+3)} \leq 1/2$. Specifically, we obtain

$$\begin{aligned} \log(N(4L(W+1)^L 2^{-b}, \mathcal{R}(d,W,L,1), L^p([0,1]^d))) & \geq c_4 W^2 L \log\left(\frac{(W+1)^L}{4L(W+1)^L 2^{-b}}\right) \\ & = c_4 W^2 L \log\left(\frac{1}{L2^{-b+2}}\right), \end{aligned} \qquad (367)$$

with $c_4 \in \mathbb{R}_+$ an absolute constant. We note that

$$\log\left(\frac{1}{L2^{-b+2}}\right) = \frac{b}{4} + \left(\frac{3b}{4} - 2 - \log(L)\right) \tag{368}$$

$$> \frac{b}{4} + \left(\frac{3}{4}(\log(L) + L\log(W+1) + 3) - 2 - \log(L)\right) \tag{369}$$

$$> \frac{b}{4} + \left(\frac{3}{4}(2\log(L) + 3) - 2 - \log(L)\right) \tag{370}$$

$$> \frac{b}{4}, \tag{371}$$

where (369) follows from the assumption $b > \log(L) + L\log(W+1) + 3$, and in (370) we used $L\log(W+1) \geq L > \log(L)$, for $L \geq 60$. Putting (366), (367), and (368)-(371) together yields

$$\log(N(L(W+1)^L 2^{-b}, \mathcal{R}_b^1(d, W, L), L^p([0,1]^d))) \geq \frac{c_4}{4} W^2 Lb.$$

As the covering number is a non-decreasing function of the covering ball radius ε, we have

$$\log(N(\varepsilon, \mathcal{R}_b^1(d, W, L), L^p([0,1]^d))) \geq \frac{c_4}{4} W^2 Lb, \quad \text{for all } \varepsilon \in (0, L(W+1)^L 2^{-b}]. \tag{372}$$

Upon setting $c_1 = \frac{c_4}{4}$, this finalizes the proof of (361).

We proceed to the proof of (362). Arbitrarily fix $\varepsilon \in (L(W+1)^L 2^{-b}, \frac{1}{32})$ and let

$$\widetilde{b} := \left\lfloor \log\left(\frac{L(W+1)^L}{2\varepsilon}\right) \right\rfloor. \tag{373}$$

We first note that

$$\widetilde{b} \leq \left\lfloor \log\left(\frac{L(W+1)^L}{2L(W+1)^L 2^{-b}}\right) \right\rfloor = \lfloor \log(2^{b-1}) \rfloor \leq b, \tag{374}$$

which leads to the inclusion $\mathcal{R}_{\widetilde{b}}^1(d, W, L) \subseteq \mathcal{R}_b^1(d, W, L)$. Thanks to Lemma F.1, we hence get

$$\log(N(\varepsilon, \mathcal{R}_b^1(d, W, L), L^p([0,1]^d))) \geq \log(N(2\varepsilon, \mathcal{R}_{\widetilde{b}}^1(d, W, L), L^p([0,1]^d))). \tag{375}$$

To lower-bound the right-hand-side of (375), we apply (372) with b replaced by $\widetilde{b}$, ε replaced by 2ε, the prerequisite $\widetilde{b} > \log(L) + L\log(W+1) + 3$ satisfied owing to

$$\widetilde{b} > \log(L) + L\log(W+1) + \log\left(\frac{1}{2\varepsilon}\right) - 1 \tag{376}$$

$$> \log(L) + L\log(W+1) + \log\left(\frac{1}{2 \cdot \frac{1}{32}}\right) - 1 \tag{377}$$

$$= \log(L) + L\log(W+1) + 3, \tag{378}$$

and the prerequisite $2\varepsilon \in (0, L(W+1)^L 2^{-\widetilde{b}}]$ satisfied thanks to

$$2\varepsilon = L(W+1)^L 2^{-\log\left(\frac{L(W+1)^L}{2\varepsilon}\right)} \leq L(W+1)^L 2^{-\widetilde{b}}, \tag{379}$$

to obtain

$$\log(N(2\varepsilon, \mathcal{R}_{\widetilde{b}}^1(d, W, L), L^p([0,1]^d))) \geq \frac{c_4}{4} W^2 L\widetilde{b}. \tag{380}$$

We further note that

$$\widetilde{b} = \left\lfloor \log\left(\frac{L(W+1)^L}{2\varepsilon}\right) \right\rfloor \tag{381}$$

$$\geq \log\left(\frac{L(W+1)^L}{2\varepsilon}\right) - 1 \tag{382}$$

$$\geq \frac{1}{2}\log\left(\frac{L(W+1)^L}{2\varepsilon}\right), \tag{383}$$

$$> \frac{1}{2}\log\left(\frac{(W+1)^L}{\varepsilon}\right), \tag{384}$$

where in (383) we used $\log(\frac{L(W+1)^L}{2\varepsilon}) \geq \log(\frac{60(61)^{61}}{2\frac{1}{32}}) > 2$ as $W, L \geq 60$ and $\varepsilon < \frac{1}{32}$, both by assumption, and (384) follows from $L \geq 60$. Using the lower bound (381)-(384) in (380), results in $\log(N(2\varepsilon, \mathcal{R}_{\widetilde{b}}^1(d,W,L), L^p([0,1]^d))) \geq \frac{c_4}{8}W^2 L \log(\frac{(W+1)^L}{\varepsilon})$, which together with (375) yields

$$\log(N(\varepsilon, \mathcal{R}_b^1(d,W,L), L^p([0,1]^d))) \geq \frac{c_4}{8}W^2 L \log\left(\frac{(W+1)^L}{\varepsilon}\right). \tag{385}$$

As ε was chosen arbitrarily in $(L(W+1)^L 2^{-b}, \frac{1}{32})$, (385) concludes the proof of (362) upon setting $c_2 = \frac{c_4}{8}$.

It remains to establish (363). To this end, we first consider the case $L(W+1)^L 2^{-b} \geq \frac{1}{32}$. It follows from (361) that

$$\log(N(\varepsilon, \mathcal{R}_b^1(d,W,L), L^p([0,1]^d))) \geq c_1 W^2 L b, \quad \text{for } \varepsilon \in (0, 1/32] \subseteq (0, L(W+1)^L 2^{-b}].$$

For $L(W+1)^L 2^{-b} < \frac{1}{32}$, we obtain from (361) and (362) that

$$\log(N(\varepsilon, \mathcal{R}_b^1(d,W,L), L^p([0,1]^d))) \geq \begin{cases} c_1 W^2 L b, & \text{for } \varepsilon \in (0, L(W+1)^L 2^{-b}], \\ c_2 W^2 L \log\left(\frac{(W+1)^L}{\varepsilon}\right), & \text{for } \varepsilon \in \left(L(W+1)^L 2^{-b}, \frac{1}{32}\right). \end{cases}$$

Combining these results, we obtain, for $\varepsilon \in (0, \frac{1}{32})$,

$$\log(N(\varepsilon, \mathcal{R}_b^1(d,W,L), L^p([0,1]^d))) \geq \min\left\{c_1 W^2 L b, c_2 W^2 L \log\left(\frac{(W+1)^L}{\varepsilon}\right)\right\}$$

$$\geq \min\{c_1, c_2\} W^2 L \cdot \min\left\{b, \log\left(\frac{(W+1)^L}{\varepsilon}\right)\right\},$$

which, upon setting $c_3 = \min\{c_1, c_2\}$, concludes the proof. $\qquad\square$

We continue with the case of moderate b. A key ingredient here is the depth-precision tradeoff developed in [18], formalized as follows.

Lemma E.3. *Let $d, W, L, k \in \mathbb{N}$. For all $a, b \in \mathbb{N}$, it holds that*

$$\mathcal{R}_{kb}^{ka}(d,W,L) \subseteq \mathcal{R}_b^a(d, 16W, (k+2)L). \tag{386}$$

Proof. The case $d = 1$ is [18, Proposition 4.1]. The proof for general d is structurally identical to that of [18, Proposition 4.1] and is provided, for completeness, in Appendix E.1. $\qquad\square$

The lower bound on the covering number of $\mathcal{R}_b^1(d,W,L)$, for moderate values of b, is as follows.

Lemma E.4. *Let $p \in [1, \infty]$, $d, W, L, b \in \mathbb{N}$, with $W, L \geq 960$. Assume that $72000 \frac{\log(W)}{L} < b \leq 4L \log(W)$. Then, we have*

$$\log(N(\varepsilon, \mathcal{R}_b^1(d, W, L), L^p([0,1]^d))) \geq c\, W^2 L b, \quad \text{for all } \varepsilon \in \left(0, \frac{1}{100}\right), \tag{387}$$

with $c \in \mathbb{R}_+$ an absolute constant.

Proof. Let $k \in \left\{2, \ldots, \lfloor \frac{L}{60} \rfloor - 2\right\}$ be an integer to be determined later. Application of Lemma E.3 with $a = 1$, W replaced by $\lfloor W/16 \rfloor$, and L replaced by $\lfloor L/(k+2) \rfloor$, yields

$$\mathcal{R}_{kb}^k(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor) \subseteq \mathcal{R}_b^1(d, 16 \lfloor W/16 \rfloor, (k+2) \lfloor L/(k+2) \rfloor). \tag{388}$$

As $k \geq 2$, we have the inclusion relation

$$\mathcal{R}_{kb}^1(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor) \subseteq \mathcal{R}_{kb}^k(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor). \tag{389}$$

Noting that $16 \lfloor W/16 \rfloor \leq W$ and $(k+2) \lfloor L/(k+2) \rfloor \leq L$, it follows that

$$\mathcal{R}_b^1(d, 16 \lfloor W/16 \rfloor, (k+2) \lfloor L/(k+2) \rfloor) \subseteq \mathcal{R}_b^1(d, W, L). \tag{390}$$

Combining (389) and (390) with (388), we obtain

$$\mathcal{R}_{kb}^1(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor) \subseteq \mathcal{R}_b^1(d, W, L). \tag{391}$$

Application of (441) in Lemma F.1 to the inclusion relation (391) then yields, for all $\varepsilon \in \mathbb{R}_+$,

$$N(\varepsilon, \mathcal{R}_b^1(d, W, L), L^p([0,1]^d)) \geq N(2\varepsilon, \mathcal{R}_{kb}^1(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor), L^p([0,1]^d)). \tag{392}$$

We next lower-bound the covering number of $\mathcal{R}_{kb}^1(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor)$ to get a lower bound on the covering number of $\mathcal{R}_b^1(d, W, L)$. To this end, we identify a value of $k \in \left\{2, \ldots, \lfloor \frac{L}{60} \rfloor - 2\right\}$ that allows us to apply Lemma E.2 to $\mathcal{R}_{kb}^1(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor)$. This is done as follows. Define $f : \mathbb{R}_+ \mapsto \mathbb{R}$ as $f(x) = xb - 5 \frac{L}{x} \log(W)$, and note that f is strictly increasing on its domain and satisfies

$$f(1) = b - 5L \log(W) < 0, \tag{393}$$

$$f\left(\left\lfloor \frac{L}{60} \right\rfloor - 2\right) > f\left(\frac{L}{120}\right) \tag{394}$$

$$= \frac{L}{120} b - 5 \frac{L}{\frac{L}{120}} \log(W) \tag{395}$$

$$= \frac{1}{120} (Lb - 72000 \log(W)) \tag{396}$$

$$> 0. \tag{397}$$

Here, in (393) we used the assumption $b \leq 4L \log(W)$, (394) follows from $\lfloor \frac{L}{60} \rfloor - 2 > \frac{L}{60} - 3 \geq \frac{L}{60} - \frac{1}{320} L > \frac{L}{120}$, as $L \geq 960$, and in (397) we invoked the assumption $b > 72000 \frac{\log(W)}{L}$. We now choose $k \in \left\{2, \ldots, \lfloor \frac{L}{60} \rfloor - 2\right\}$ to be the unique integer such that $f(k-1) < 0$ and $f(k) \geq 0$, namely,

$$(k-1)b - 5 \frac{L}{k-1} \log(W) < 0 \tag{398}$$

$$kb - 5 \frac{L}{k} \log(W) \geq 0. \tag{399}$$

For this k, we now lower-bound the covering number of $\mathcal{R}^1_{kb}(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor)$ by applying Lemma E.2 with b replaced by kb, W replaced by $\lfloor W/16 \rfloor$, and L replaced by $\lfloor L/(k+2) \rfloor$. We first verify the prerequisites of Lemma E.2 by noting that $\lfloor W/16 \rfloor \geq \lfloor 960/16 \rfloor = 60$, $\lfloor L/(k+2) \rfloor \geq \lfloor L/(\lfloor \frac{L}{60} \rfloor - 2 + 2) \rfloor \geq 60$, and

$$kb \geq 5\frac{L}{k}\log(W) \tag{400}$$

$$\geq \frac{L}{k}\log(W) + \frac{L}{k}\log(W) + 3\frac{L}{\lfloor \frac{L}{60} \rfloor - 2}\log(W) \tag{401}$$

$$\geq \log(\lfloor L/(k+2) \rfloor) + \lfloor L/(k+2) \rfloor \log(\lfloor W/16 \rfloor + 1) + 3, \tag{402}$$

where (401) follows from $k \in \{2, \ldots, \lfloor \frac{L}{60} \rfloor - 2\}$, and in (402) we used $\frac{L}{k}\log(W) \geq \lfloor L/(k+2) \rfloor \geq \log(\lfloor L/(k+2) \rfloor)$, $\frac{L}{k}\log(W) \geq \lfloor L/(k+2) \rfloor \log(\lfloor W/16 \rfloor + 1)$, and $\frac{L}{\lfloor \frac{L}{60} \rfloor - 2}\log(W) \geq 1$. Then, application of (363) in Lemma E.2 with $\varepsilon = \frac{1}{50}$, yields

$$\log(N(1/50, \mathcal{R}^1_{kb}(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor), L^p([0,1]^d))) \tag{403}$$

$$\geq c_3(\lfloor W/16 \rfloor)^2 \lfloor L/(k+2) \rfloor) \cdot \min\left\{ kb, \log\left(\frac{(\lfloor W/16 \rfloor + 1)^{\lfloor L/(k+2) \rfloor}}{1/50}\right)\right\}, \tag{404}$$

with $c_3 \in \mathbb{R}_+$ an absolute constant. To lower-bound the right-hand-side of (404), we note that

$$\lfloor W/16 \rfloor \geq \frac{1}{2} \cdot \frac{W}{16} = \frac{W}{32}, \tag{405}$$

$$\lfloor L/(k+2) \rfloor \geq \lfloor L/(4k) \rfloor \geq \frac{L}{8k}, \tag{406}$$

and

$$\log\left(\frac{(\lfloor W/16 \rfloor + 1)^{\lfloor L/(k+2) \rfloor}}{1/50}\right) \geq \lfloor L/(k+2) \rfloor \log(\lfloor W/16 \rfloor + 1) \tag{407}$$

$$\geq \lfloor L/(k+2) \rfloor \log(W/16) \tag{408}$$

$$\geq \frac{1}{16}\frac{L}{k-1}\log(W) \tag{409}$$

$$> \frac{1}{80}(k-1)b \tag{410}$$

$$\geq \frac{1}{160}kb, \tag{411}$$

where in (409) we used $\lfloor L/(k+2) \rfloor \geq \frac{1}{2} \cdot L/(k+2) \geq \frac{1}{8}\frac{L}{k-1}$ and $\log(W/16) \geq \frac{1}{2}\log(W)$, (410) follows from (398), and (411) is thanks to $k \geq 2$. Then, using (405), (406), and (407)-(411) in (403)-(404), yields

$$\log(N(1/50, \mathcal{R}^1_{kb}(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor), L^p([0,1]^d))) \tag{412}$$

$$\geq c_3\left(\frac{W}{32}\right)^2\left(\frac{L}{8k}\right)\min\left\{kb, \frac{1}{160}kb\right\} \tag{413}$$

$$= \frac{c_3}{32^2 \cdot 8 \cdot 160}W^2 Lb. \tag{414}$$

As the covering number is non-decreasing in ε, we have shown that, for all $\varepsilon \in (0, \frac{1}{100})$,

$$\log(N(2\varepsilon, \mathcal{R}^1_{kb}(d, \lfloor W/16 \rfloor, \lfloor L/(k+2) \rfloor), L^p([0,1]^d))) \geq \frac{c_3}{32^2 \cdot 8 \cdot 160}W^2 Lb. \tag{415}$$

We conclude the proof by putting (392) and (415) together and setting $c = \frac{c_3}{32^2 \cdot 8 \cdot 160}$. $\qquad \square$

We are now ready to proceed to the proof of the lower bound (140). Let $D := 960$ and $E := 2 \cdot 10^5$ and arbitrarily fix an $\varepsilon \in (0, \frac{1}{100})$. By assumption, we hence have $W \geq D \geq 960$, $L \geq D \geq 960$, and

$$L(a + b) \geq E \log(W) = 2 \cdot 10^5 \log(W). \tag{416}$$

It follows from (352) in Lemma E.1 that

$$N\big(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d)\big) = N\bigg(\frac{\varepsilon}{2^{(a-1)L}}, R_{a+b-1}^1(d, W, L), L^p([0,1]^d)\bigg). \tag{417}$$

We proceed by distinguishing two cases, the first one being

$$(a + b - 1) > \log(L) + L\log(W + 1) + 3. \tag{418}$$

Then, it follows from (363) in Lemma E.2 with ε replaced by $\frac{\varepsilon}{2^{(a-1)L}}$, b replaced by $a + b - 1$, and the prerequisite $(a + b - 1) > \log(L) + L\log(W + 1) + 3$ satisfied thanks to (418), that, for $\frac{\varepsilon}{2^{(a-1)L}} \in (0, \frac{1}{100})$,

$$\log\bigg(N\bigg(\frac{\varepsilon}{2^{(a-1)L}}, \mathcal{R}_{a+b-1}^1(d, W, L), L^p([0,1]^d)\bigg)\bigg) \tag{419}$$

$$\geq c_1 W^2 L \cdot \min\bigg\{(a + b - 1), \log\bigg(\frac{(W+1)^L 2^{(a-1)L}}{\varepsilon}\bigg)\bigg\}, \tag{420}$$

with $c_1 \in \mathbb{R}_+$ an absolute constant. The second case is

$$(a + b - 1) \leq \log(L) + L\log(W + 1) + 3. \tag{421}$$

Using $a + b - 1 \geq \frac{1}{2}(a + b)$ in (416) and $\log(L) + L\log(W+1) + 3 < L + L\log(W^2) + L < 4L\log(W)$ in (421), yields

$$72000\frac{\log(W)}{L} < 100000\frac{\log(W)}{L} \leq (a + b - 1) < 4L\log(W). \tag{422}$$

Now, application of Lemma E.4 with ε replaced by $\frac{\varepsilon}{2^{(a-1)L}}$, b replaced by $a + b - 1$, the prerequisite $72000\frac{\log(W)}{L} < a + b - 1 \leq 4L\log(W)$ satisfied thanks to (422), yields, for $\frac{\varepsilon}{2^{(a-1)L}} \in (0, \frac{1}{100})$,

$$\log\bigg(N\bigg(\frac{\varepsilon}{2^{(a-1)L}}, \mathcal{R}_{a+b-1}^1(d, W, L), L^p([0,1]^d)\bigg)\bigg) \tag{423}$$

$$\geq c_2 W^2 L(a + b - 1) \tag{424}$$

$$\geq c_2 W^2 L \cdot \min\bigg\{(a + b - 1), \log\bigg(\frac{(W+1)^L 2^{(a-1)L}}{\varepsilon}\bigg)\bigg\}, \tag{425}$$

with $c_2 \in \mathbb{R}_+$ an absolute constant. Combining (419)-(420) and (423)-(425), yields, for all $a, b \in \mathbb{N}$,

$$\log\bigg(N\bigg(\frac{\varepsilon}{2^{(a-1)L}}, \mathcal{R}_{a+b-1}^1(d, W, L), L^p([0,1]^d)\bigg)\bigg)$$
$$\geq \min\{c_1, c_2\}W^2 L \cdot \min\bigg\{(a + b - 1), \log\bigg(\frac{(W+1)^L 2^{(a-1)L}}{\varepsilon}\bigg)\bigg\}. \tag{426}$$

Using

$$(a + b - 1) \geq \frac{1}{2}(a + b), \tag{427}$$

$$\log\left(\frac{(W+1)^L 2^{(a-1)L}}{\varepsilon}\right) = \frac{1}{2}\log\left(\frac{(W+1)^{2L} 2^{2(a-1)L}}{\varepsilon^2}\right) \tag{428}$$

$$\geq \frac{1}{2}\log\left(\frac{(W+1)^{2L} 2^{(a-1)L}}{\varepsilon^2}\right) \tag{429}$$

$$\geq \frac{1}{2}\log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right), \tag{430}$$

together with (417) in (426), yields

$$\log\big(N(\varepsilon, \mathcal{R}_b^a(d, W, L), L^p([0,1]^d))\big) \geq \frac{\min\{c_1, c_2\}}{2} W^2 L \cdot \min\left\{(a + b), \log\left(\frac{(W+1)^L 2^{aL}}{\varepsilon}\right)\right\}.$$

The proof is concluded upon setting $c = \frac{\min\{c_1, c_2\}}{2}$.

E.1 Proof of Lemma E.3

We first state the following technical lemma.

Lemma E.5. *[18, Proposition F.1] Let $d, W, L \in \mathbb{N}$, and let $\mathbb{A} \subseteq \mathbb{R}$ be a finite set satisfying $\{-1, 0, 1\} \subseteq \mathbb{A}$. Then, for every $k \in \mathbb{N}$ and all $u, v \in \mathbb{A} \cap \mathbb{R}_{\geq 0}$, it holds that*

$$\mathcal{R}_{\mathcal{T}_1(\mathbb{A}, u, v, k)}(d, W, L) \subseteq \mathcal{R}_{\mathbb{A}}(d, 16W, (k+3)L)$$

with

$$\mathcal{T}_1(\mathbb{A}, u, v, k) := \left\{\sum_{i=0}^{k}(u^i \alpha_i + v^i \beta_i) : |\alpha_i|, |\beta_i| \in \mathbb{A}, i = 0, \ldots, k\right\}. \tag{431}$$

We are now ready to prove Lemma E.3. For $k = 1$, (386) is trivially satisfied. For $k \geq 2$, we note that

$$\mathcal{T}_1(\mathbb{Q}_b^a, 2^{-b}, 2^a, k - 1) \tag{432}$$

$$= \left\{\sum_{i=0}^{k-1}(2^{-bi}\alpha_i + 2^{ai}\beta_i) : |\alpha_i|, |\beta_i| \in \mathbb{Q}_b^a, i = 0, \ldots, k-1\right\} \tag{433}$$

$$= \left\{\sum_{i=0}^{k-1}(2^{-bi}\alpha_i + 2^{ai}\beta_i) : \alpha_i, \beta_i \in \mathbb{Q}_b^a, i = 0, \ldots, k-1\right\} \tag{434}$$

$$\supseteq \left\{\pm\sum_{i=-kb}^{ka} 2^i c_i : c_i \in \{0, 1\}\right\} \tag{435}$$

$$= \mathbb{Q}_{kb}^{ka}, \tag{436}$$

where in (435) we used $\mathbb{Q}_b^a = \{\pm\sum_{i=-b}^{a}\theta_i 2^i : \theta_i \in \{0, 1\}\}$. Thanks to (432)-(436), we have

$$\mathcal{R}_{kb}^{ka}(d, W, L) \subseteq \mathcal{R}_{\mathcal{T}_1(\mathbb{Q}_b^a, 2^{-b}, 2^a, k-1)}(d, W, L). \tag{437}$$

Application of Lemma E.5 with $u = 2^{-b}$, $v = 2^a$, $\mathbb{A} = \mathbb{Q}_b^a$, and k replaced by $k - 1$, yields

$$\mathcal{R}_{\mathcal{T}_1(\mathbb{Q}_b^a, 2^{-b}, 2^a, k-1)}(d, W, L) \subseteq \mathcal{R}_b^a(d, 16W, (k+2)L). \tag{438}$$

The proof is finalized by combining (437) and (438) to obtain (386).

F Auxiliary results

F.1 Relation Between Covering Number and Packing Number

The following lemma on covering and packing numbers is frequently used throughout this paper.

Lemma F.1. *Let $(\mathcal{X}, \delta)$ be a metric space and $\varepsilon \in \mathbb{R}_+$. It holds that*

$$M(2\varepsilon, \mathcal{X}, \delta) \leq N(\varepsilon, \mathcal{X}, \delta) \leq M(\varepsilon, \mathcal{X}, \delta). \tag{439}$$

Let $\mathcal{Y} \subseteq \mathcal{X}$. We have,

$$M(\varepsilon, \mathcal{X}, \delta) \geq M(\varepsilon, \mathcal{Y}, \delta), \tag{440}$$

$$N(\varepsilon, \mathcal{X}, \delta) \geq N(2\varepsilon, \mathcal{Y}, \delta), \tag{441}$$

Proof. Relation (439) is [6, Lemma 5.5]. To prove (440), we simply note that every ε-packing of $\mathcal{Y}$ is also an ε-packing of $\mathcal{X}$, and hence

$$M(\varepsilon, \mathcal{X}, \delta) \geq M(\varepsilon, \mathcal{Y}, \delta). \tag{442}$$

Then, (441) follows from $N(\varepsilon, \mathcal{X}, \delta) \geq M(2\varepsilon, \mathcal{X}, \delta) \geq M(2\varepsilon, \mathcal{Y}, \delta) \geq N(2\varepsilon, \mathcal{X}, \delta)$, where the first and the last inequalities are by (439) and the second by (442). $\qquad\square$

F.2 Augmenting Networks

This section is concerned with a technical lemma, which shows how a ReLU network of a given depth can be augmented to a deeper network while retaining its input-output relation.

Lemma F.2. *Let $d, W, L \in \mathbb{N}$, with $W \geq 2$, and let $\mathbb{A} \subseteq \mathbb{R}$, with $\{-1, 0, 1\} \subseteq \mathbb{A}$. For every $f \in \mathcal{R}_{\mathbb{A}}(d, W, L)$, there exists a network configuration $\Phi \in \mathcal{N}_{\mathbb{A}}(d, W, L)$ such that $\mathcal{L}(\Phi) = L$ and $\mathcal{R}(\Phi) = f$.*

A special case of Lemma F.2, namely $\mathbb{A} = \mathbb{R}$, is documented in [18, Lemma H.2]. The proof of Lemma F.2 is almost identical to that of [18, Lemma H.2], but will be provided for completeness.

Proof of Lemma F.2. By definition, there exists a network configuration $\widetilde{\Phi} = ((\widetilde{A}_\ell, \widetilde{b}_\ell))_{\ell=1}^{\widetilde{L}} \in \mathcal{N}_{\mathbb{A}}(d, W, L)$, with $\widetilde{L} \leq L$, such that $R(\widetilde{\Phi}) = f$. If $\widetilde{L} = L$, setting $\Phi = \widetilde{\Phi}$, we have $\mathcal{L}(\Phi) = L$ and $\mathcal{R}(\Phi) = f$. For $L > \widetilde{L}$, we let $\Phi := ((A_\ell, b_\ell))_{\ell=1}^{L}$, with

$$(A_\ell, b_\ell) := (\widetilde{A}_\ell, \widetilde{b}_\ell), \quad \text{for } 1 \leq \ell < \widetilde{L},^{10} \tag{443}$$

$A_{\widetilde{L}} := \begin{pmatrix} \widetilde{A}_{\widetilde{L}} \\ -\widetilde{A}_{\widetilde{L}} \end{pmatrix}$, $b_{\widetilde{L}} := \begin{pmatrix} \widetilde{b}_{\widetilde{L}} \\ -\widetilde{b}_{\widetilde{L}} \end{pmatrix}$, $A_\ell := I_2$, $b_\ell := 0_2$, for ℓ such that $\widetilde{L} < \ell < L$, and $A_L := (1 \;\; -1)$, $b_L := 0$. Invoking the assumption $\{-1, 0, 1\} \subseteq \mathbb{A}$, this yields $\Phi \in \mathcal{N}_{\mathbb{A}}(d, W, L)$, with $\mathcal{L}(\Phi) = L$. We now note that

$$S(A_L, b_L) \circ \rho \circ \cdots \circ \rho \circ S(A_{\widetilde{L}}, b_{\widetilde{L}}) = S(A_L, b_L) \circ \rho \circ S(A_{\widetilde{L}}, b_{\widetilde{L}}) \tag{444}$$

$$= S(\widetilde{A}_{\widetilde{L}}, \widetilde{b}_{\widetilde{L}}), \tag{445}$$

where in (444) we used $\rho \circ S(A_\ell, b_\ell) = \rho \circ S(I_2, 0_2) = \rho$, for $\widetilde{L} < \ell < L$, and $\rho \circ \rho = \rho$, and (445) is by $\left(S(A_L, b_L) \circ \rho \circ S(A_{\widetilde{L}}, b_{\widetilde{L}}) \right)(x) = \rho(\widetilde{A}_{\widetilde{L}} x + \widetilde{b}_{\widetilde{L}}) - \rho(-\widetilde{A}_{\widetilde{L}} x - \widetilde{b}_{\widetilde{L}}) = \widetilde{A}_{\widetilde{L}} x + \widetilde{b}_{\widetilde{L}} = S(\widetilde{A}_{\widetilde{L}}, \widetilde{b}_{\widetilde{L}})(x)$, for $x \in \mathbb{R}_{d'}$, with d' denoting the number of columns of $\widetilde{A}_{\widetilde{L}}$. Combining (444)-(445) and (443), we see that $R(\Phi) = R(\widetilde{\Phi}) = f$. $\qquad\square$

---

[10]Here and in what follows, we use the convention that if there does not exist an ℓ, in this case, satisfying the constraint, the assignment is skipped; in the present case, this would apply if $\widetilde{L} = 1$.

F.3 Existence of the Empirical Risk Minimizer in (83)

Arbitrarily fix a sample $(x_i, y_i)_{i=1}^n \in ([0,1] \times \mathbb{R})^n$. The existence of the empirical risk minimizer, for this fixed sample, is equivalent to the existence of a network configuration $\Phi \in \mathcal{N}(1, \lceil D+1 \rceil, L(n), 1)$ whose associated truncated realization $\mathcal{T}_1(R(\Phi))$ minimizes the empirical risk. Noting that the set $\mathcal{N}(1, \lceil D+1 \rceil, L(n), 1)$ can be written as a finite disjoint union of network configurations, each element in the union corresponding to a given network architecture, we only have to show the existence of a minimizer over each given element in this union. Arbitrarily fix an element in the union with associated architecture $(N_0, \ldots, N_\ell)$ and let $\mathcal{N}_{N_0,\ldots,N_\ell}$ be the corresponding set of network configurations. Next, note that $\mathcal{N}_{N_0,\ldots,N_\ell} = [-1,1]^{\sum_{i=1}^\ell (N_i N_{i-1} + N_i)}$ is a compact set and, by Lemma 2.2, the mapping $\Phi \in \mathcal{N}_{N_0,\ldots,N_\ell} \mapsto \frac{1}{n} \sum_{i=1}^n (\mathcal{T}_1(R(\Phi)(x_i)) - y_i)^2$ is continuous. As continuous functions on compact sets attain minima, there exists a $\Phi \in \mathcal{N}_{N_0,\ldots,N_\ell}$ that minimizes the empirical risk within the set $\mathcal{N}_{N_0,\ldots,N_\ell}$, as was to be shown. The argument is concluded by noting that the choice of the sample $(x_i, y_i)_{i=1}^n \in ([0,1] \times \mathbb{R})^n$ was arbitrary and hence there exists an empirical risk minimizer for each sample $(x_i, y_i)_{i=1}^n \in ([0,1] \times \mathbb{R})^n$.

References

[1] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan 1989.

[2] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, Dec 1989.

[3] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, Jan 1989.

[4] A. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[5] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, no. 2, pp. 264–280, 1971.

[6] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, 2nd ed. Cambridge University Press, 2019, vol. 48.

[7] Z. Shen, H. Yang, and S. Zhang, "Deep network approximation characterized by number of neurons," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1768–1811, 2020.

[8] M. Kohler and S. Langer, "On the rate of convergence of fully connected deep neural network regression estimates," *The Annals of Statistics*, vol. 49, no. 4, pp. 2231 – 2249, 2021.

[9] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks." *Journal of Machine Learning Research*, vol. 20, no. 63, pp. 1–17, 2019.

[10] D. Elbrächter, D. Perekrestenko, P. Grohs, and H. Bölcskei, "Deep neural network approximation theory," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2581–2623, May 2021.

[11] D. Yarotsky and A. Zhevnerchuk, "The phase diagram of approximation rates for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 13 005–13 015.

[12] A. Caragea, D. G. Lee, J. Maly, G. Pfander, and F. Voigtlaender, "Quantitative approximation results for complex-valued neural networks," *SIAM Journal on Mathematics of Data Science*, vol. 4, no. 2, pp. 553–580, 2022.

[13] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, vol. 48, no. 4, pp. 1875 – 1897, 2020.

[14] M. Chen, H. Jiang, W. Liao, and T. Zhao, "Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery," *Information and Inference: A Journal of the IMA*, vol. 11, no. 4, pp. 1203–1253, Mar. 2022.

[15] A. Caragea, P. Petersen, and F. Voigtlaender, "Neural network approximation and estimation of classifiers with classification boundary in a Barron class," *The Annals of Applied Probability*, vol. 33, no. 4, pp. 3039 – 3079, 2023.

[16] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999, vol. 9.

[17] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] W. Ou, P. Schenkel, and H. Bölcskei, "Three quantization regimes for ReLU networks," *arXiv preprint arXiv:2405.01952*, 2024.

[19] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab, "DNN expression rate analysis of high-dimensional PDEs: Application to option pricing," *Constructive Approximation*, vol. 55, no. 1, pp. 3–71, 2022.

[20] C. Huyen, *Designing Machine Learning Systems*. O'Reilly Media, Inc., 2022.

[21] R. Nakada and M. Imaizumi, "Adaptive approximation and generalization of deep neural network with intrinsic dimensionality," *Journal of Machine Learning Research*, vol. 21, no. 174, pp. 1–38, 2020.

[22] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural Networks*, vol. 94, pp. 103 – 114, 2017.

[23] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 8–45, 2019.

[24] I. Gühring and M. Raslan, "Approximation rates for neural networks with encodable weights in smoothness spaces," *Neural Networks*, vol. 134, pp. 107–130, 2021.

[25] P. Petersen and F. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep ReLU neural networks," *Neural Networks*, vol. 108, pp. 296–330, 2018.

[26] S. A. Janowsky, "Pruning versus clipping in neural networks," *Physical Review A*, vol. 39, no. 12, pp. 6600–6603, June 1989.

[27] D. Blalock, J. J. Gonzalez Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" *Proceedings of Machine Learning and Systems*, vol. 2, pp. 129–146, 2020.

[28] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[29] J. Maly and R. Saab, "A simple approach for quantizing neural networks," *Applied and Computational Harmonic Analysis*, vol. 66, pp. 138–150, 2023.

[30] G. Vardi, G. Yehudai, and O. Shamir, "Width is less important than depth in ReLU neural networks," in *Proc. of Thirty Fifth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 178.   PMLR, July 2022, pp. 1249–1281.

[31] M. Telgarsky, "Benefits of depth in neural networks," in *29th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 49, Columbia University, New York, USA, 23–26 June 2016, pp. 1517–1539.

[32] B. M. Sh and M. Solomjak, *Quantitative analysis in Sobolev imbedding theorems and applications to spectral theory*, ser. American Mathematical Society translations. American Mathematical Society, 1980.

[33] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, vol. 10, no. 4, pp. 1040 – 1053, Dec 1982.

[34] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A distribution-free theory of nonparametric regression*, ser. Springer Series in Statistics.   Springer, 2002, vol. 1.

[35] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, vol. 27, no. 4, pp. 1564–1599, 1999.

[36] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Applied and Computational Harmonic Analysis*, vol. 1, no. 1, pp. 100–115, 1993.

[37] ——, "Unconditional bases and bit-level compression," *Applied and Computational Harmonic Analysis*, vol. 3, no. 4, pp. 388–392, 1996.

[38] P. Grohs, "Optimally sparse data representations," in *Harmonic and Applied Analysis: From Groups to Signals*.   Springer, 2015, pp. 199–248.

[39] Z. Shen, H. Yang, and S. Zhang, "Nonlinear approximation via compositions," *Neural Networks*, vol. 119, pp. 74–84, 2019.

[40] S. Mendelson and R. Vershynin, "Entropy and the combinatorial dimension," *Inventiones Mathematicae*, vol. 152, no. 1, pp. 37–55, 2003.