Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ_1 and ℓ_0 controls

Article · July 2016

CITATIONS
6

READS
154

2 authors:

Jason M. Klusowski
Rutgers, The State University of New Jersey
12 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Andrew R Barron
Yale University
93 PUBLICATIONS **8,783** CITATIONS
SEE PROFILE

Approximation by Combinations of ReLU and Squared ReLU Ridge Functions with ℓ^1 and ℓ^0 Controls

Jason M. Klusowski Andrew R. Barron *†

May 23, 2018

Abstract

We establish L^∞ and L^2 error bounds for functions of many variables that are approximated by linear combinations of ReLU (rectified linear unit) and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls on their inner and outer parameters. With the squared ReLU ridge function, we show that the L^2 approximation error is inversely proportional to the inner layer ℓ^0 sparsity and it need only be sublinear in the outer layer ℓ^0 sparsity. Our constructions are obtained using a variant of the Jones-Barron probabilistic method, which can be interpreted as either stratified sampling with proportionate allocation or two-stage cluster sampling. We also provide companion error lower bounds that reveal near optimality of our constructions. Despite the sparsity assumptions, we showcase the richness and flexibility of these ridge combinations by defining a large family of functions, in terms of certain spectral conditions, that are particularly well approximated by them.

1 Introduction

Functions of many variables are approximated using linear combinations of ridge functions with one layer of nonlinearities, viz.,

$$f_m(x) = \sum_{k=1}^{m} b_k \phi(a_k \cdot x - t_k), \tag{1}$$

where $b_k \in \mathbb{R}$ are the outer layer parameters and $a_k \in \mathbb{R}^d$ are the vectors of inner parameters for the single-hidden layer of functions $\phi(a_k \cdot x - t_k)$. The activation function ϕ is allowed to be quite general. For example, it can be bounded and

*Jason M. Klusowski and Andrew R. Barron are with the Department of Statistics and Data Science, Yale University, New Haven, CT, USA, 06511, e-mail: {jason.klusowski, andrew.barron}@yale.edu.

†

Lipschitz, polynomials with certain controls on their degrees, or bounded with jump discontinuities. When the ridge activation function is a sigmoid, (1) is single-hidden layer artificial neural network.

One goal in a statistical setting is to estimate a regression function, i.e., conditional mean response, $f(x) = \mathbb{E}[Y \mid X = x]$ with domain $D \triangleq [-1,1]^d$ from noisy observations $\{(X_i, Y_i)\}_{i=1}^n$, where $Y = f(X) + \varepsilon$. In classical literature [3], $L^2(P)$ mean squared prediction error of order $(d/n)^{1/2}$, achieved by ℓ^1 penalized least squares estimators[1] over the class of models (1), are obtained by optimizing the tradeoff between *approximation error* and *descriptive complexity relative to sample size*. Bounds on the approximation error are obtained by first showing how models of the form (1) with $\phi(z) = \mathbf{1}\{z > 0\}$ can be used to approximate f satisfying $\int_{\mathbb{R}^d} \|\omega\|_1 |\mathcal{F}(f)(\omega)| d\omega < +\infty$, provided f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix\cdot\omega} \mathcal{F}(f)(\omega) d\omega$ on $[-1,1]^d$. Because it is often difficult to work with discontinuous ϕ (i.e., vanishing or exploding gradient issues), these step functions are replaced with smooth ϕ such that $\phi(\tau z) \wedge 1 \to \mathbf{1}\{z > 0\}$ as $\tau \to +\infty$. Thus, this setup allows one to work with approximants of the form (1) with smooth ϕ, but at the expense of *unbounded* ℓ^1 norm $\|a_k\|_1$.

Like high-dimensional linear regression [18], many applications of statistical inference and estimation require a setting where $d \gg n$. In contrast to the aforementioned mean square prediction error of $(d/n)^{1/2}$, it has been shown [12] how models of the form (1) with Lipschitz[2] ϕ (reps. Lipschitz derivative ϕ') and *bounded* inner parameters $\|a_k\|_0$ and $\|a_k\|_1$ can be used to give desirable $L^2(D)$ mean squared prediction error of order $((\log d)/n)^{1/3}$ (resp. $((\log d)/n)^{2/5}$), also achieved by penalized estimators.[3] In fact, [13] shows that these rates are nearly optimal. A few natural questions arise from restricting the ℓ^0 and ℓ^1 norms of the inner parameters in the model:

- To what degree do the sparsity assumptions limit the flexibility of the model (1)?
- What condition can be imposed on f so that it can be approximated by f_m with Lipschitz ϕ (or Lipschitz derivative ϕ') and bounded $\|a_k\|_0$ and / or $\|a_k\|_1$?
- How well can f be approximated by f_m, given these sparsity constraints?

According to classic approximation results [1, 2], if the domain of f is contained in $[-1,1]^d$ and f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix\cdot\omega} \mathcal{F}(f)(\omega) d\omega$, then the spectral condition $v_{f,1} < +\infty$, where $v_{f,s} \triangleq \int_{\mathbb{R}^d} \|\omega\|_1^s |\mathcal{F}(f)(\omega)| d\omega$, is enough to ensure that $f - f(0)$ can be approximated in

[1] That is, the fit minimizes $(1/n)\sum_{i=1}^n (f_m(X_i) - Y_i)^2 + \lambda \sum_{k=1}^m |b_k|$ for some appropriately chosen $\lambda > 0$.

[2] Henceforth, when we say a function is Lipschitz, we assume it has bounded Lipschitz parameter.

[3] With additional ℓ^0 inner sparsity, we might also consider an estimator that minimizes $(1/n)\sum_{i=1}^n (f_m(X_i) - Y_i)^2 + \lambda_0 \psi\left(\sum_{k=1}^m |b_k| \|a_k\|_0\right)$ for some convex function ψ and appropriately chosen $\lambda_0 > 0$.

$L^\infty(D)$ by equally weighted, i.e, $|b_1| = \cdots = |b_m|$, linear combinations of functions of the form (1) with $\phi(z) = \mathbf{1}\{z > 0\}$. Typical L^∞ error rates $\|f - f_m\|_\infty$ of an m-term approximation (1) are at most $cv_{f,1}\sqrt{d}\, m^{-1/2}$, where c is a universal constant [1, 8, 21]. A rate of $c(p)v_{f,1}m^{-1/2-1/(pd)}$ was given in [15, Theorem 3] for $L^p(D)$ for nonnegative even integer p. Again, all these bounds are valid when the step activation function is replaced by a smooth approximant ϕ (in particular, *any* sigmoid satisfying $\lim_{z\to\pm\infty}\phi(z) = \pm 1$), but at the expense of unbounded $\|a_k\|_1$.

Towards giving partial answers to the questions we posed, in Section 2, we show how functions of the form (1) with ReLU (also known as a ramp or first order spline) $\phi(z) = (z)_+ = 0 \vee z$ (which is Lipchitz)[4] or squared ReLU $\phi(z) = (z)_+^2$ (which has Lipschitz derivative) activation function can be used to give desirable $L^\infty(D)$ approximation error bounds, even when $\|a_k\|_1 = 1$, $0 \le t_k \le 1$, and $|b_1| = \cdots = |b_m|$. Because of the widespread popularity of the ReLU activation function and its variants, these simpler forms may also be of independent interest for computational and algorithmic reasons as in [4, 7, 10, 11, 22], to name a few.

Unlike the case with step activation functions, our analysis makes no use of the combinatorial properties of half-spaces as in Vapnik-Chervonenkis theory [9, 20]. The $L^2(D)$ case for ReLU ridge functions (also known as hinging hyperplanes) with ℓ^1-bounded inner parameters was considered in [6, Theorem 3] and our $L^\infty(D)$ bounds improve upon that line of work and, in addition, increase the exponent from $1/2$ to $1/2 + O(1/d)$. Our proof techniques are substantively different than [6] and, importantly, are more amenable to empirical process theory, which is the key to showing our error bounds.

These tighter rates of approximation, with ReLU and squared ReLU activation functions, are possible under two different conditions – finite $v_{f,2}$ or $v_{f,3}$, respectively. The main idea we use originates from [15] and [16] and can be seen as stratified sampling with proportionate allocation. This technique is widely applied in survey sampling as a means of variance reduction [17].

At the end of Section 2, we will also discuss the degree to which these bounds can be improved by providing companion lower bounds on the minimax rates of approximation.

Section 3 will focus on how accurate estimation can be achieved even when $\|a_k\|_0$ is also bounded. In particular, we show how an m-term linear combination (1) with $\|a_k\|_0 \le \sqrt{m}$ and $\|a_k\|_1 = 1$ can approximate f satisfying $v_{f,3} < +\infty$ in $L^2(D)$ with error at most $\sqrt{2}v_{f,3}m^{-1/2}$. In other words, the $L^2(D)$ approximation error is inversely proportional to the inner layer sparsity and it need only be sublinear in the outer layer sparsity. The constructions that achieve these error bounds are obtained using a variant of the Jones-Barron probabilistic method, which can be interpreted as two-stage cluster sampling.

Throughout this paper, we will state explicitly how our bounds depend on d so that the reader can fully appreciate the complexity of approximation. If

[4]It is perhaps more conventional to write $(z)^+$ for $0 \vee z$, however, to avoid clutter in the exponent, we use the current notation.

a is a vector in Euclidean space, we use the notation $a(k)$ to denote its k-th component.

2 L^∞ approximation with bounded ℓ^1 norm

2.1 Positive results

In this section, we provide the statements and proofs of the existence results for f_m with bounded ℓ^1 norm of inner parameters. We would like to point out that the results of Theorem 1 hold when all occurrences of the ReLU or squared ReLU activation functions are replaced by general ϕ which is Lipschitz or has Lipschitz derivative ϕ', respectively.

Theorem 1. *Suppose f admits an integral representation*

$$f(x) = v \int_{[0,1]\times\{a:\|a\|_1=1\}} \eta(t,a)\ (a\cdot x - t)_+^{s-1} dP(t,a), \tag{2}$$

for x in $D = [-1,1]^d$ and $s \in \{2,3\}$, where P is a probability measure on $[0,1]\times\{a\in\mathbb{R}^d : \|a\|_1 = 1\}$ and $\eta(t,a)$ is either -1 or $+1$. There exists a linear combination of ridge functions of the form

$$f_m(x) = \frac{v}{m}\sum_{k=1}^{m} b_k (a_k\cdot x - t_k)_+^{s-1}, \tag{3}$$

with $b_k \in [-1,1]$, $\|a_k\|_1 = 1$, $0 \le t_k \le 1$ such that

$$\sup_{x\in D} |f(x) - f_m(x)| \le c\sqrt{d+\log m}\ m^{-1/2-1/d}, \quad s = 2,$$

and

$$\sup_{x\in D} |f(x) - f_m(x)| \le c\sqrt{d}\ m^{-1/2-1/d}, \quad s = 3,$$

for some universal constant $c > 0$. Furthermore, if the b_k are restricted to $\{-1,1\}$, the upper bound is of order

$$\sqrt{d+\log m}\ m^{-1/2-1/(d+2)}, \quad s = 2$$

and

$$\sqrt{d}\ m^{-1/2-1/(d+2)}, \quad s = 3.$$

Theorem 2. *Let $D = [-1,1]^d$. Suppose f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix\cdot\omega}\mathcal{F}(f)(\omega)d\omega$ and*

$$v_{f,2} = \int_{\mathbb{R}^d} \|\omega\|_1^2 |\mathcal{F}(f)(\omega)| d\omega < +\infty.$$

There exists a linear combination of ReLU ridge functions of the form

$$f_m(x) = b_0 + a_0 \cdot x + \frac{v}{m} \sum_{k=1}^{m} b_k (a_k \cdot x - t_k)_+ \tag{4}$$

with $b_k \in [-1, 1]$, $\|a_k\|_1 = 1$, $0 \le t_k \le 1$, $b_0 = f(0)$, $a_0 = \nabla f(0)$, *and* $v \le 2v_{f,2}$ *such that*

$$\sup_{x \in D} |f(x) - f_m(x)| \le c v_{f,2} \sqrt{d + \log m}\ m^{-1/2-1/d},$$

for some universal constant $c > 0$. *Furthermore, if the* b_k *are restricted to* $\{-1, 1\}$, *the upper bound is of order*

$$v_{f,2} \sqrt{d + \log m}\ m^{-1/2-1/(d+2)}.$$

Theorem 3. *Under the setup of Theorem 2, suppose*

$$v_{f,3} = \int_{\mathbb{R}^d} \|\omega\|_1^3 |\mathcal{F}(f)(\omega)| d\omega < +\infty.$$

There exists a linear combination of squared ReLU ridge functions of the form

$$f_m(x) = b_0 + a_0 \cdot x + x^T A_0 x + \frac{v}{2m} \sum_{k=1}^{m} b_k (a_k \cdot x - t_k)_+^2 \tag{5}$$

with $b_k \in [-1, 1]$, $\|a_k\|_1 = 1$, $0 \le t_k \le 1$, $b_0 = f(0)$, $a_0 = \nabla f(0)$, $A_0 = \nabla \nabla^T f(0)$, *and* $v \le 2v_{f,3}$ *such that*

$$\sup_{x \in D} |f(x) - f_m(x)| \le c v_{f,3} \sqrt{d}\ m^{-1/2-1/d},$$

for some universal constant $c > 0$. *Furthermore, if the* b_k *are restricted to* $\{-1, 1\}$, *the upper bound is of order*

$$v_{f,3} \sqrt{d}\ m^{-1/2-1/(d+2)}.$$

The key observation for proving Theorem 2 and Theorem 3 is that f modulo linear or quadratic terms with finite $v_{f,s}$ can be written in the integral form (2). Unlike in [6, Theorem 3] where an interpolation argument is used, our technique of writing f as the mean of a random variable allows for more straightforward use of empirical process theory to bound the expected sup-error of the empirical average of m independent draws from its population mean. Our argument is also more flexible than [6] and can be readily adapted to the case of squared ReLU activation function. We should also point out that our $L^\infty(D)$ error bounds immediately imply $L^p(D)$ error bounds for all $p \ge 1$. In fact, using nearly exactly the same techniques, it can be shown that the results in Theorem 1, Theorem 2, and Theorem 3 hold verbatim in $L^2(D)$, sans the $\sqrt{d + \log m}$ or $\sqrt{d}$ factors, corresponding to the ReLu or squared ReLU cases, respectively.

Remark 1. *In [16], it was shown that the standard order $m^{-1/2}$ $L^\infty(D)$ error bound alluded to earlier could be improved to be of order $\sqrt{\log m}\ m^{-1/2-1/(2d)}$ under an alternate condition of finite $v^\star_{f,1} \triangleq \sup_{u\in\mathbb{S}^{d-1}} \int_0^\infty r^d |\mathcal{F}(f)(ru)| dr$, but with the requirement that $\|a_k\|_1$ be unbounded. In general, our assumptions are neither stronger nor weaker than this since the function f with Fourier transform $\mathcal{F}(f)(\omega) = e^{-\|\omega-\omega_0\|}/\|\omega-\omega_0\|$ for $\omega_0 \neq 0$ and $d \geq 2$ has infinite $v^\star_{f,1}$ but finite $v_{f,s}$ for $s \geq 0$, while the function f with Fourier transform $\mathcal{F}(f)(\omega) = 1/(1+\|\omega\|)^{d+2}$ has finite $v^\star_{f,1}$ but infinite $v_{f,s}$ for $s \geq 2$.*

Proof of Theorem 1. **Case I: $s = 2$.** Let $\mathcal{B}_1, \ldots, \mathcal{B}_M$ be a partition of the space $\Omega = \{(\eta, t, a)' : \eta \in \{-1, +1\},\ 0 \leq t \leq 1,\ \|a\|_1 = 1\}$ such that

$$\inf_{(\widetilde{\eta},\widetilde{t},\widetilde{a})'\in\mathcal{B}_k,\ k=1,\ldots,M} \sup_{(\eta,t,a)'\in\Omega} \|h(\widetilde{\eta},\widetilde{t},\widetilde{a}) - h(\eta,t,a)\|_\infty < \epsilon, \tag{6}$$

where $h(\eta, t, a)(x) = h(x) = \eta(a\cdot x - t)_+^{s-1}$. It is not hard to show that $M \asymp \epsilon^{-d}$. For $k = 1, \ldots, M$ define

$$dP_k(t,a) = dP(t,a)\mathbf{1}\{(\eta(t,a), t, a)' \in \mathcal{B}_k\}/L_k,$$

where L_k is chosen to make P_k a probability measure. A very important property we will use is that $\mathsf{Var}_{P_k}[h] \leq \epsilon$, which follows from (6). Let m be a positive integer and define a sequence of M independent random variables $\{m_k\}_{1\leq k\leq M}$ as follows: let m_k equal $\lfloor mL_k \rfloor$ and $\lceil mL_k \rceil$ with probabilities chosen to make its mean equal to mL_k. Given, $\underline{m} = \{m_k\}_{1\leq k\leq M}$, take a random sample $\underline{a} = \{(t_{j,k}, a_{j,k})'\}_{1\leq j\leq n_k,\ 1\leq k\leq M}$ of size $n_k = m_k + \mathbf{1}\{m_k = 0\}$ from P_k. Thus, we split the population Ω into M "strata" $\mathcal{B}_1, \ldots, \mathcal{B}_M$ and allocate the number of within-stratum samples to be proportional to the "size" of the stratum $m_1, \ldots, m_M$ (i.e., proportionate allocation). The within-stratum variability of h (i.e., $\mathsf{Var}_{P_k}[h]$) is now smaller than the population level variability (i.e., $\mathsf{Var}_P[h]$) by a factor of ϵ as evidenced by (6).

Note that the n_k sum to be at most $m + M$ because

$$\begin{aligned}
\sum_{k=1}^M n_k &= \sum_{k=1}^M m_k \mathbf{1}\{m_k > 0\} + \sum_{k=1}^M \mathbf{1}\{m_k = 0\} \\
&\leq \sum_{k=1}^M (mL_k + 1)\mathbf{1}\{m_k > 0\} + \sum_{k=1}^M \mathbf{1}\{m_k = 0\} \\
&= m\sum_{k=1}^M L_k \mathbf{1}\{m_k > 0\} + M \\
&\leq m + M,
\end{aligned} \tag{7}$$

where the last inequality follows from $\sum_{k=1}^M L_k \leq 1$. For $j = 1, \ldots, m_k$, let $h_{j,k} = h(\eta(t_{j,k}, a_{j,k}), t_{j,k}, a_{j,k})$ and $f_k = \frac{vm_k}{mn_k}\sum_{j=1}^{n_k} h_{j,k}$. Also, let $\overline{f}_m = \sum_{k=1}^M f_k$. A simple calculation shows that the mean of $\overline{f}_m$

is f. Write $\sum_{k=1}^{M}(f_k(x) - \mathbb{E}f_k(x)) = \frac{v}{m}\left(\sum_{k=1}^{M}(m_k - L_k m)\mathbb{E}_{P_k}h(x)\right) + \frac{v}{m}\left(\sum_{k=1}^{M}\sum_{j=1}^{n_k}\frac{m_k}{n_k}(h_{j,k}(x) - \mathbb{E}_{P_k}h(x))\right)$. By the triangle inequality, we upper bound

$$\mathbb{E}\sup_{x\in D}|\overline{f}_m(x) - f(x)| = \mathbb{E}\sup_{x\in D}|\sum_{k=1}^{M}(f_k(x) - \mathbb{E}f_k(x))|$$

by

$$\frac{v}{m}\mathbb{E}_{\underline{m}}\sup_{x\in D}|\sum_{k=1}^{M}(m_k - L_k m)\mathbb{E}_{P_k}h(x)| +$$

$$\frac{v}{m}\mathbb{E}_{\underline{m}}\mathbb{E}_{\underline{a}|\underline{m}}\sup_{x\in D}|\sum_{k=1}^{M}\sum_{j=1}^{n_k}\frac{m_k}{n_k}(h_{j,k}(x) - \mathbb{E}_{P_k}h(x))|. \tag{8}$$

Now

$$\mathbb{E}_{\underline{a}|\underline{m}}\sup_{x\in D}|\sum_{k=1}^{M}\sum_{j=1}^{n_k}\frac{m_k}{n_k}(h_{j,k}(x) - \mathbb{E}_{P_k}h(x))| \leq$$

$$2\mathbb{E}_{\underline{a}|\underline{m}}\sup_{x\in D}|\sum_{k=1}^{M}\sum_{j=1}^{n_k}\sigma_{j,k}\frac{m_k}{n_k}[h_{j,k}(x) - \mu_{j,k}(x)]|, \tag{9}$$

where $\{\sigma_{j,k}\}$ is a sequence of independent identically distributed Rademacher variables and $\{x \mapsto \mu_{j,k}(x)\}$ is any sequence of functions defined on D [see for example Lemma 2.3.6 in [19]]. For notational brevity, we define $\widetilde{h}_{j,k}(x) = \frac{m_k}{n_k}[h_{j,k}(x) - \mu_{j,k}(x)]$. By Dudley's entropy integral method [see Corollary 13.2 in [5]], the quantity in (9) can be bounded by

$$24\int_0^{\delta/2}\sqrt{N(u,D)}du, \tag{10}$$

where $N(u,D)$ is the u-metric entropy of D with respect to the norm $\kappa(x,x')$ (i.e., the logarithm of the smallest u-net that covers D with respect to κ) defined by

$$\kappa^2(x,x') \triangleq \sum_{k=1}^{M}\sum_{j=1}^{n_k}(\widetilde{h}_{j,k}(x) - \widetilde{h}_{j,k}(x'))^2$$
$$\leq (m+M)\|x - x'\|_\infty^2, \tag{11}$$

and $\delta^2 = \sup_{x\in D}\sum_{k=1}^{M}\sum_{j=1}^{n_k}|\widetilde{h}_{j,k}(x)|^2$. If we set $\mu_{j,k}$ to equal $\frac{m_k}{n_k}h(\eta(t_k,a_k),t_k,a_k)$, where $(\eta_k,t_k,a_k)'$ is any fixed point in $\mathcal{B}_k$, it follows from (6) and (7) that $\delta \leq \sqrt{m+M}\epsilon$ and from (11) that $N(u,D) \leq d\log(3\sqrt{m+M}/u)$. By evaluating the integral in (10), we can bound the second term in (8) by

$$24v\sqrt{d}\ m^{-1/2}\epsilon\sqrt{-\log\epsilon + 1}\sqrt{1 + M/m}. \tag{12}$$

For the first expectation in (8), we follow a similar approach. As before,

$$\begin{aligned}&\mathbb{E}_{\underline{m}} \sup_{x\in D} |\sum_{k=1}^{M}(m_k - L_k m)\mathbb{E}_{P_k} h(x)| \\ &\leq 2\mathbb{E}_{\underline{m}} \sup_{x\in D} |\sum_{k=1}^{M}\sigma_k(m_k - L_k m)\mathbb{E}_{P_k} h(x)|, \end{aligned} \tag{13}$$

where $\{\sigma_k\}$ is a sequence of independent identically distributed Rademacher variables. For notational brevity, we write $\widetilde{h}_k(x) = (m_k - L_k m)\mathbb{E}_{P_k} h(x)$. We can also bound (13) by (10), except this time $N(u, D)$ is the u-metric entropy of D with respect to the norm $\rho(x, x')$ defined by

$$\begin{aligned}\rho^2(x, x') &\triangleq \sum_{k=1}^{M}(\widetilde{h}_k(x) - \widetilde{h}_k(x'))^2 \\ &\leq M\|x - x'\|_\infty^2, \end{aligned} \tag{14}$$

where the last line follows from $|m_k - L_k m| \leq 1$ and $|\mathbb{E}_{P_k} h(x) - \mathbb{E}_{P_k} h(x')| \leq \|x - x'\|_\infty$. The quantity δ is also less than $\sqrt{M}$, since $\sup_{x\in D} |\widetilde{h}_k(x)| \leq 1$ and moreover $N(u, D) \leq d \log(3\sqrt{M}/u)$. Evaluating the integral in (10) with these specifications yields a bound on the first term in (8) of

$$\frac{48v\sqrt{d}\sqrt{M}}{m}. \tag{15}$$

Adding (15) and (12) together yields a bound on $\mathbb{E}\sup_{x\in D} |\overline{f}_m(x) - f(x)|$ of

$$48v\sqrt{d}m^{-1/2}(\sqrt{M/m} + \epsilon\sqrt{1 + M/m}\sqrt{-\log\epsilon + 1}). \tag{16}$$

Choose

$$M = m\frac{\epsilon^2(-\log\epsilon + 1)}{1 - \epsilon^2(-\log\epsilon + 1)}. \tag{17}$$

Consequently, $\mathbb{E}\sup_{x\in D} |\overline{f}_m(x) - f(x)|$ is at most

$$96v\sqrt{d}m^{-1/2}\frac{\epsilon\sqrt{-\log\epsilon + 1}}{\sqrt{1 - \epsilon^2(-\log\epsilon + 1)}}. \tag{18}$$

We stated earlier that $M \asymp \epsilon^{-d}$. Thus (17) determines ϵ to be at most of order $m^{-1/(d+2)}$. Since the inequality (17) holds on average, there is a realization of $\overline{f}_m$ for which $\sup_{x\in D} |\overline{f}_m(x) - f(x)|$ has the same bound. Note that $\overline{f}_m$ has the desired equally weighted form.

For the second conclusion, we set $m_k = mL_k$ and $n_k = \lceil m_k \rceil$. In this case, the first term in (8) is zero and hence $\mathbb{E}\sup_{x\in D} |\overline{f}_m(x) - f(x)|$ is not greater than (12). The conclusion follows with $M = m$ and ϵ of order $m^{-1/d}$.

Case II: $s = 3$. The metric $\kappa(x, x')$ is in fact bounded by a constant multiple of $\sqrt{m + M}\epsilon\|x - x'\|_\infty$. To see this, we note that the function $\widetilde{h}_{j,k}(x)$ has the form

$$\pm\frac{m_k}{n_k}[(a \cdot x - t)_+^2 - (a_k \cdot x - t_k)_+^2],$$

with $\|a - a_k\|_1 + |t - t_k| < \epsilon$. Thus, the gradient of $\widetilde{h}_{j,k}(x)$ with respect to x has the form

$$\nabla\widetilde{h}_{j,k}(x) = \pm\frac{2m_k}{n_k}[(a(a \cdot x - t)_+ - a_k(a_k \cdot x - t_k)_+].$$

Adding and subtracting $\frac{2m_k}{n_k}a(a_k \cdot x - t_k)_+$ to the above expression yields the bound of order ϵ for $\sup_{x\in D}\|\nabla\widetilde{h}_{j,k}(x)\|_1$. Taylor's theorem yields the desired bound on $\kappa(x, x')$. Again using Dudley's entropy integral, we can bound $\mathbb{E}\sup_{x\in D}|\overline{f}_m(x) - f(x)|$ by a universal constant multiple of either $v\sqrt{d}m^{-1/2}(\sqrt{M/m}+\epsilon\sqrt{1 + M/m})$ or $v\sqrt{d}m^{-1/2}\epsilon\sqrt{1 + M/m}$ corresponding to the equally weighted or non-equally weighted cases, respectively. The corresponding results follow with $M = m\epsilon^2/(1 - \epsilon^2)$ and ϵ of order $m^{-1/(d+2)}$ or $M = m$ and ϵ of order $m^{-1/d}$. Note that here the additional smoothness afforded by the stronger assumption $v_{f,3} < +\infty$ allows one to remove the $\sqrt{-\log\epsilon + 1}$ factor that appeared in the final bound in the proof of Theorem 2. This rate is the same as what was achieved in Theorem 2, without a $\sqrt{(\log m)/d + 1}$ factor. □

Proof of Theorem 2. If $|z| \le c$, we note the identity

$$-\int_0^c[(z - u)_+e^{iu} + (-z - u)_+e^{-iu}]du = e^{iz} - iz - 1. \tag{19}$$

If $c = \|\omega\|_1$, $z = \omega \cdot x$, $a = a(\omega) = \omega/\|\omega\|_1$, and $u = \|\omega\|_1 t$, $0 \le t \le 1$, we find that

$$-\|\omega\|_1^2\int_0^1[(a \cdot x - t)_+e^{i\|\omega\|_1 t} + (-a \cdot x - t)_+e^{-i\|\omega\|_1 t}]dt = e^{i\omega\cdot x} - i\omega \cdot x - 1.$$

Multiplying the above by $\mathcal{F}(f)(\omega) = e^{ib(\omega)}|\mathcal{F}(f)(\omega)|$, integrating over $\mathbb{R}^d$, and applying Fubini's theorem yields

$$f(x) - x \cdot \nabla f(0) - f(0) = \int_{\mathbb{R}^d}\int_0^1 g(t, \omega)dtd\omega,$$

where

$$g(t, \omega) = -[(a \cdot x - t)_+\cos(\|\omega\|_1 t + b(\omega)) + (-a \cdot x - t)_+\cos(\|\omega\|_1 t - b(\omega))]\|\omega\|_1^2|\mathcal{F}(f)(\omega)|.$$

Consider the probability measure on $\{-1,1\} \times [0,1] \times \mathbb{R}^d$ defined by

$$dP(z,t,\omega) = \frac{1}{v} |\cos(z\|\omega\|_1 t + b(\omega))| \|\omega\|_1^2 |\mathcal{F}(f)(\omega)| dt d\omega, \tag{20}$$

where

$$v = \int_{\mathbb{R}^d} \int_0^1 [|\cos(\|\omega\|_1 t + b(\omega))| + |\cos(\|\omega\|_1 t - b(\omega))|] \|\omega\|_1^2 |\mathcal{F}(f)(\omega)| dt d\omega \leq 2v_{f,2}.$$

Define a function $h(z,t,a)(x)$ that equals

$$(za \cdot x - t)_+ \ \eta(z,t,\omega),$$

where $\eta(z,t,\omega) = -\text{sgn}\cos(\|\omega\|_1 zt + b(\omega))$. Note that $h(z,t,a)(x)$ has the form $\pm(\pm a \cdot x - t)_+$. Thus, we see that

$$f(x) - x \cdot \nabla f(0) - f(0) = v \int_{\{-1,1\} \times [0,1] \times \mathbb{R}^d} h(z,t,a)(x) dP(z,t,\omega). \tag{21}$$

The result follows from an application of Theorem 1. □

Proof of Theorem 3. For the result in Theorem 3, we will use exactly the same techniques. The function $f(x) - x^T \nabla \nabla^T f(0) x/2 - x \cdot \nabla f(0) - f(0)$ can be written as the real part of

$$\int_{\mathbb{R}^d} (e^{i\omega \cdot x} + (\omega \cdot x)^2/2 - i\omega \cdot x - 1) \mathcal{F}(f)(\omega) d\omega. \tag{22}$$

As before, the integrand in (22) admits an integral representation given by

$$(i/2)\|\omega\|_1^3 \int_0^1 [(-a \cdot x - t)_+^2 e^{-i\|\omega\|_1 t} - (a \cdot x - t)_+^2 e^{i\|\omega\|_1 t}] dt,$$

which can be used to show that $f(x) - x^T \nabla \nabla^T f(0) x/2 - x \cdot \nabla f(0) - f(0)$ equals

$$\frac{v}{2} \int_{\{-1,1\} \times [0,1] \times \mathbb{R}^d} h(z,t,a)(x) dP(z,t,\omega), \tag{23}$$

where

$$h(z,t,a) = \text{sgn}\sin(z\|\omega\|_1 t + b(\omega)) \ (za \cdot x - t)_+^2$$

and

$$dP(z,t,\omega) = \frac{1}{v} |\sin(z\|\omega\|_1 t + b(\omega))| \|\omega\|_1^3 |\mathcal{F}(f)(\omega)| dt d\omega,$$

$$v = \int_{\mathbb{R}^d} \int_0^1 [|\sin(\|\omega\|_1 t + b(\omega))| + |\sin(\|\omega\|_1 t - b(\omega))|] \|\omega\|_1^3 |\mathcal{F}(f)(\omega)| dt d\omega \leq 2v_{f,3}.$$

The result follows from an application of Theorem 1. □

Remark 2. *By slightly modifying the definition of h from the proofs of Theorem 2 and Theorem 3 (in particular, multiplying it by a sinusoidal function of ω and t), it suffices to sample instead from the density $dP(t,\omega) = \frac{\|\omega\|_1^s |\mathcal{F}(f)(\omega)|}{v_{f,s}} dt d\omega$ on $[0,1] \times \mathbb{R}^d$.*

Remark 3. *For unit bounded x, the expression $e^{i\omega\cdot x} - i\omega \cdot x - 1$ is bounded in magnitude by $\|\omega\|_1^2$, so one only needs Fourier representation of $f(x) - x \cdot \nabla f(0) - f(0)$ when using the integrability with the $\|\omega\|_1^2$ factor. Similarly, $e^{i\omega\cdot x} + (\omega\cdot x)^2/2 - i\omega\cdot x - 1$ is bounded in magnitude by $\|\omega\|^3$, so one only needs Fourier representation of $f(x) - x^T \nabla\nabla^T f(0) x - x\cdot\nabla f(0) - 1$ when using the integrability with the $\|\omega\|_1^3$ factor.*

Remark 4. *Note that in Theorem 2 and Theorem 3, we work with integrals with respect to the absolutely continuous measure $d\mathcal{F}(f)(\omega)$. In general, a (complex) Fourier measure $d\mathcal{F}(f)(\omega)$ does not need to be absolutely continuous. For instance, it can be discrete on a lattice of values of ω, associated with a multivariate Fourier series representation for bounded domains x (and periodic extensions thereof). Indeed, for bounded domains, one might have access to both Fourier series and Fourier transforms of extensions of f to $\mathbb{R}^d$. The best extension is one that gives the smallest Fourier norm $\int_{\mathbb{R}^d} \|\omega\|_1^s |d\mathcal{F}(f)(\omega)|$. For further discussion along these lines, see [2].*

Next, we investigate the optimality of the rates from Section 2.

2.2 Lower bounds

Let $\mathcal{H}_s = \{x \mapsto \eta(a\cdot x - t)_+^{s-1} : \|a\|_1 \le 1,\ 0 \le t \le 1,\ \eta \in \{-1,+1\}\}$ and for $p \in [2,+\infty]$ let $\mathcal{F}_p^s$ denote the closure of the convex hull of $\mathcal{H}_s$ with respect to the $\|\cdot\|_p$ norm on $L^p(D,P)$ for p finite, where P is the uniform probability measure on D, and $\|\cdot\|_\infty$ (the supremum norm over D) for $p = +\infty$. We let $\mathcal{C}_m^s$ denote the collection of all convex combinations of m terms from $\mathcal{H}_s$. By Theorem 2 and Theorem 3, after possibly subtracting a linear or quadratic term, $f/(2v_{f,2})$ and $f/v_{f,3}$ belongs to $\mathcal{F}_p^2$ and $\mathcal{F}_p^3$, respectively. For $p \in [2,+\infty]$ and $\epsilon > 0$, we define the ϵ-covering number $N_p(\epsilon)$ by

$$\min\{n : \exists\ \mathcal{F} \subset \mathcal{F}_p^s,\ |\mathcal{F}| = n,\ \text{s.t.}\ \inf_{f'\in\mathcal{F}} \sup_{f\in\mathcal{F}_p^s} \|f - f'\|_p \le \epsilon\}.$$

and the ϵ-packing number $M_p(\epsilon)$ by

$$\max\{n : \exists\ \mathcal{F} \subset \mathcal{F}_p^s,\ |\mathcal{F}| = n,\ \text{s.t.}\ \inf_{f,f'\in\mathcal{F}} \|f - f'\|_p > \epsilon\}.$$

Theorem 1 implies that $\inf_{f_m\in\mathcal{C}_m^s} \sup_{f\in\mathcal{F}_\infty^s} \|f - f_m\|_\infty$ achieves the bounds as stated therein.

Theorem 4. *For $p \in [2,+\infty]$ and $s \in \{2,3\}$,*

$$\inf_{f_m\in\mathcal{C}_m^s} \sup_{f\in\mathcal{F}_p^s} \|f - f_m\|_p \ge (Amd^{2s+1}\log(md))^{-1/2-s/d},$$

for some universal positive constant A.

Ignoring the dependence on d and logarithmic factors in m, this result coupled with Theorem 1 implies that $\inf_{f_m \in \mathcal{C}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_p$ is between $m^{-1/2-2/d}$ and $m^{-1/2-1/d}$; for large d, the rates are essentially the same. Compare this with [15, Theorem 4] or [1, Theorem 3], where a lower bound of $c(\delta, d)\ m^{-1/2-1/d-\delta}$, $\delta > 0$ arbitrary, was obtained for approximants of the form (1) with Lipschitz ϕ, but with inner parameter vectors of unbounded ℓ^1 norm.

We only give the proof of Theorem 4 for $s = 2$, since the other case $s = 3$ is handled similarly. First, we provide a few ancillary results that will be used later on. The next result is contained in [14, Lemma 4.2] and is useful for giving a lower bound on $M_p(\epsilon)$.

Lemma 1. *Let H be a Hilbert space equipped with a norm $\|\cdot\|$ and containing a finite set $\mathcal{H}$ with the following properties.*

(i) $|\mathcal{H}| \geq 3$,

(ii) $\sum_{h,h' \in \mathcal{H},\ h \neq h'} |\langle h, h' \rangle| \leq \delta^2$

(iii) $\delta^2 \leq \min_{h \in \mathcal{H}} \|h\|^2$

Then there exists a collection $\Omega \subset \{0,1\}^{|\mathcal{H}|}$ with cardinality at least $2^{(1-H(1/4))|\mathcal{H}|-1}$, where $H(1/4)$ is the entropy of a Bernoulli random variable with success probability $1/4$, such that each pair of elements in the set $\mathcal{F} = \left\{ \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \omega_h h : (\omega_h : h \in \mathcal{H}) \in \Omega \right\}$ is separated by at least $\frac{1}{2}\sqrt{\frac{\min_{h \in \mathcal{H}} \|h\|^2 - \delta^2}{|\mathcal{H}|}}$ in $\|\cdot\|$.

Lemma 2. *If θ belongs to $[R]^d = \{1, 2, \ldots, R\}^d$, $R \in \mathbb{Z}^+$, then the collection of functions*

$$\mathcal{H} = \{x \mapsto \sin(\pi\theta \cdot x)/(4\pi\|\theta\|_1^2) : \theta \in [R]^d\}$$

satisfies the assumption of Lemma 1 with $H = L^2(D, P)$, where P is the uniform probability measure on D. Moreover, $|\mathcal{H}| = R^d$, $\delta = 0$, $\min_{h \in \mathcal{H}} \|h\| = 1/(4\sqrt{2}\pi d^2 R^2)$, and $\mathcal{F} \subset \mathcal{F}_p^1$ for all $p \in [2, +\infty]$. Consequently, if $\epsilon = 1/(8\sqrt{2}\pi d^2 R^{2+d/2})$, then

$$\begin{aligned} \log M_p(\epsilon) &\geq (\log 2)(1 - H(1/4)) \left(8\epsilon\sqrt{2}\pi d^2\right)^{-\frac{2d}{4+d}} - 1 \\ &\geq \left(c\epsilon d^2\right)^{-\frac{2d}{4+d}}, \end{aligned} \tag{24}$$

for some universal constant $c > 0$.

Proof. We first observe the identity

$$\begin{aligned} \sin(\pi\theta \cdot x)/(4\pi\|\theta\|_1^2) &= \theta \cdot x/(4\pi\|\theta\|_1^2) + \\ &\frac{\pi}{4} \int_0^1 [(-a \cdot x - t)_+ - (a \cdot x - t)_+] \sin(\pi\|\theta\|_1 t) dt, \end{aligned}$$

where $a = a(\theta) = \theta/\|\theta\|_1$. Note that above integral can also be written as an expectation of

$$-z \operatorname{sgn}(\sin(\pi\|\theta\|_1 t))\ (za \cdot x - t)_+ \in \mathcal{H}_2$$

with respect to the density

$$p_\theta(z,t) = \frac{\pi}{4}|\sin(\pi\|\theta\|_1 t)|,$$

on $\{-1,1\} \times [0,1]$. The fact that p_θ integrates to one is a consequence of the identity

$$\int_0^1 |\sin(\pi\|\theta\|_1 t)|dt = 2/\pi.$$

Since $\int_D |\sin(\pi\theta \cdot x)|^2 dP(x) = 1/2$, each member of $\mathcal{H}$ has norm equal to $1/(4\sqrt{2}\pi\|\theta\|_1^2)$ and each pair of elements is orthogonal so that $\delta = 0$. Integrations over D involving $\sin(\pi\theta \cdot x)$ are easiest to see using an instance of Euler's formula, viz., $\sin(\alpha \cdot x) = \frac{1}{2i}(\prod_{k=1}^d e^{i\alpha(k)x(k)} - \prod_{k=1}^d e^{-i\alpha(k)x(k)})$. □

Proof of Theorem 4. Let $A > 0$ be arbitrary. Suppose contrary to the hypothesis,

$$\inf_{f_m \in \mathcal{C}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_p < (Amd^5 \log(md))^{-1/2-2/d}$$
$$\triangleq \epsilon_0/3.$$

Note that each element of $\mathcal{C}_m^2$ has the form $\sum_{k=1}^m \lambda_k h_k$, where $\sum_{k=1}^m \lambda_k = 1$ and $h_k \in \mathcal{H}_s$. Next, consider the subcollection $\widetilde{\mathcal{C}}_m^2$ with elements of the form $\sum_{k=1}^m \widetilde{\lambda}_k \widetilde{h}_k$, where $\widetilde{\lambda}_k$ belongs to an $\epsilon_0/3$-net $\widetilde{\mathcal{P}}$ of the $m-1$ dimensional probability simplex $\mathcal{P}_m$ and $\widetilde{h}_k$ belongs to an $\epsilon_0/3$-net $\widetilde{\mathcal{H}}$ of $\mathcal{H}_s$. By a stars and bars argument, there are at most $|\widetilde{\mathcal{P}}|\binom{m+|\mathcal{H}|-1}{m}$ such functions. Furthermore, since $\sup_{h \in \mathcal{H}_s} \|h\|_\infty \le 1$, we have

$$\begin{aligned}\inf_{f_m \in \widetilde{\mathcal{C}}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_2 \le & \inf_{f_m \in \mathcal{C}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_2 + \\ & \inf_{\widetilde{h} \in \widetilde{\mathcal{H}}} \sup_{h \in \mathcal{H}_s} \|h - \widetilde{h}\|_2 + \\ & \inf_{\widetilde{\lambda} \in \widetilde{\mathcal{P}}} \sup_{\lambda \in \mathcal{P}_m} \|\lambda - \widetilde{\lambda}\|_1 \\ & < \epsilon_0/3 + \epsilon_0/3 + \epsilon_0/3 = \epsilon_0.\end{aligned}$$

Since $|\widetilde{\mathcal{H}}| \asymp \epsilon_0^{-d-1}$ and $|\widetilde{\mathcal{P}}| \asymp \epsilon_0^{-m+1}$, it follows that

$$\begin{aligned}\log N_p(\epsilon_0) &\le \log|\widetilde{\mathcal{C}}_m^2| \\ &\le c_0 \log\left[\epsilon_0^{-m-1}\binom{m + c_1\epsilon_0^{-d-1} - 1}{m}\right] \\ &\le c_2 dm \log(1/\epsilon_0) \\ &\le c_3 dm \log(Adm), \end{aligned} \tag{25}$$

for some positive universal constants $c_0 > 0$, $c_1 > 0$, $c_2 > 0$, and $c_3 > 0$.

On the other hand, using (24) from Lemma 2 coupled with the fact that $N_p(\epsilon_0) \geq M_p(2\epsilon_0)$, we have

$$\begin{aligned} \log N_p(\epsilon_0) &\geq \log M_p(2\epsilon_0) \\ &\geq \left(2c\epsilon_0 d^2\right)^{-\frac{2d}{4+d}} \\ &\geq c_4 Adm \log(dm), \end{aligned} \tag{26}$$

for some universal constant $c_4 > 0$. Combining (25) and (26), we find that

$$c_4 Adm \log(dm) \leq c_3 dm \log(Adm).$$

If A is large enough (independent of m or d), we reach a contradiction. This proves the lower bound. □

3 L^2 approximation with bounded ℓ^0 and ℓ^1 norm

In Section 2, we explored conditions for which good approximation in $L^\infty(D)$ could be achieved even with ℓ^1 controls on the inner parameter vectors. In this section, we show how similar statements can be made in $L^2(D)$, but with control on the ℓ^0 norm as well. Note that unlike Theorem 1, we see in Theorem 5 how the smoothness of the activation function directly affects the rate of approximation. The proof is obtained by applying the Jones-Barron probabilistic method in two stages (similar to two-stage cluster sampling), first on the outer layer coefficients, and then on the inner layer coefficients.

Theorem 5. *Suppose f admits an integral representation*

$$f(x) = v \int_{[0,1]\times\{a:\|a\|_1=1\}} \eta(t,a)\ (a \cdot x - t)_+^{s-1} dP(t,a),$$

for x in $D = [-1,1]^d$ and $s \in \{2,3\}$, where P is a probability measure on $[0,1] \times \{a \in \mathbb{R}^d : \|a\|_1 = 1\}$ and $\eta(t,a)$ is either -1 or $+1$. There exists a linear combination of ridge functions of the form

$$f_{m,m_0}(x) = \frac{v}{m} \sum_{k=1}^{m} b_k \left(a_k \cdot x - t_k\right)_+^{s-1},$$

where $\|a_k\|_0 \leq m_0$, $\|a_k\|_1 = 1$, and $b_k \in \{-1,+1\}$ such that

$$\|f - f_{m,m_0}\|_2 \leq v\sqrt{\frac{1}{m} + \frac{1}{m_0^{s-1}}}.$$

Furthermore, the same rates for $s = 2$ or $s = 3$ are achieved for general f adjusted by a linear or quadratic term with $v = 2v_{f,2} < +\infty$ or $v = v_{f,3} < +\infty$, respectively.

Remark 5. *In particular, taking $m_0 = \sqrt{m}$, it follows that there exists an m-term linear combination of squared ReLU ridge functions, with $\sqrt{m}$-sparse inner parameter vectors, that approximates f with $L^2(D)$ error at most $\sqrt{2}vm^{-1/2}$. In other words, the $L^2(D)$ approximation error is inversely proportional to the inner layer sparsity and it need only be sublinear in the outer layer sparsity.*

Proof. Take a random sample $\underline{a} = \{(t_k, a_k)'\}_{1 \le k \le m}$ from P. Given $\underline{a}$, take a random sample $\underline{\widetilde{a}} = \{\widetilde{a}_{\ell,k}\}_{1 \le \ell \le m_0,\ 1 \le k \le m}$, where $\mathbb{P}[\widetilde{a}_{\ell,k} = \mathrm{sgn}(a_k(j))e_j] = |a_k(j)|$ for $j = 1, \ldots, d$, $a_k = (a_k(1), \ldots, a_k(d))'$, and e_j is the j-th standard basis vector for $\mathbb{R}^d$. Note that

$$\mathbb{E}_{\underline{\widetilde{a}}|\underline{a}}[\widetilde{a}_{\ell,k}] = a_k \tag{27}$$

and

$$\begin{aligned} \mathsf{Var}_{\underline{\widetilde{a}}|\underline{a}}[\widetilde{a}_{\ell,k} \cdot x] \le \mathbb{E}_{\underline{\widetilde{a}}|\underline{a}}[\widetilde{a}_{\ell,k} \cdot x]^2 &= \sum_{j=1}^{d} |a_k(j)||x(j)|^2 \\ &\le \|a_k\|_1 \|x\|_\infty^2 \le 1. \end{aligned} \tag{28}$$

Define

$$\overline{f}_{m,m_0}(x) = \frac{v}{m} \sum_{k=1}^{m} \eta(t_k, a_k) \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x - t_k \right)_+^{s-1}. \tag{29}$$

By the bias-variance decomposition,

$$\mathbb{E}\|f - \overline{f}_{m,m_0}\|_2^2 = \mathbb{E}\|\overline{f}_{m,m_0} - \mathbb{E}\overline{f}_{m,m_0}\|_2^2 + \|f - \mathbb{E}\overline{f}_{m,m_0}\|_2^2.$$

Note that $\mathbb{E}\|\overline{f}_{m,m_0} - \mathbb{E}\overline{f}_{m,m_0}\|_2^2 \le \frac{v^2}{m}$. Next, observe that

$$\begin{aligned} f(x) - \mathbb{E}\overline{f}_{m,m_0}(x) = \frac{v}{m} \sum_{k=1}^{m} \mathbb{E}_{\underline{a}} \Bigg[\eta(t_k, a_k) \times \\ \mathbb{E}_{\underline{\widetilde{a}}|\underline{a}} \left((a_k \cdot x - t_k)_+^{s-1} - \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x - t_k \right)_+^{s-1} \right) \Bigg], \end{aligned}$$

which, by an application of the triangle inequality, implies that

$$\begin{aligned} |f(x) - \mathbb{E}\overline{f}_{m,m_0}(x)| \le \frac{v}{m} \sum_{k=1}^{m} \\ \mathbb{E}_{\underline{a}} \left| (a_k \cdot x - t_k)_+^{s-1} - \mathbb{E}_{\underline{\widetilde{a}}|\underline{a}} \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x - t_k \right)_+^{s-1} \right|. \end{aligned}$$

Next, we use the following two properties of $(z)_+^{s-1}$: for all z and z' in $\mathbb{R}$,

$$|(z)_+ - (z')_+| \le |z - z'|, \tag{30}$$

$$|(z)_+^2 - (z')_+^2 - 2(z - z')(z')_+| \le |z - z'|^2. \tag{31}$$

If $s = 2$, we have by (30), (27), and (28) that

$$\mathbb{E}_{\underline{a}} \left| (a_k \cdot x - t_k)_+ - \mathbb{E}_{\widetilde{\underline{a}}|\underline{a}} \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x - t_k \right)_+ \right| \leq$$
$$\mathbb{E}_{\underline{a}} \mathbb{E}_{\widetilde{\underline{a}}|\underline{a}} \left| a_k \cdot x - \frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x \right| \leq$$
$$\mathbb{E}_{\underline{a}} \sqrt{\mathbb{E}_{\widetilde{\underline{a}}|\underline{a}} \left| a_k \cdot x - \frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x \right|^2} =$$
$$\mathbb{E}_{\underline{a}} \sqrt{\frac{\mathsf{Var}_{\widetilde{\underline{a}}|\underline{a}}[\widetilde{a}_{\ell,k} \cdot x]}{m_0}} \leq \frac{1}{\sqrt{m_0}}.$$

This shows that $\|f - \mathbb{E}\overline{f}_{m,m_0}\|_2^2 \leq \frac{v^2}{m_0}$. If $s = 3$, we have from (31), (27), and (28) that

$$\mathbb{E}_{\underline{a}} \left| (a_k \cdot x - t_k)_+^2 - \mathbb{E}_{\widetilde{\underline{a}}|\underline{a}} \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x - t_k \right)_+^2 \right| \leq$$
$$\mathbb{E}_{\underline{a}} \mathbb{E}_{\widetilde{\underline{a}}|\underline{a}} \left| a_k \cdot x - \frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k} \cdot x \right|^2 =$$
$$\mathbb{E}_{\underline{a}} \left[\frac{\mathsf{Var}_{\widetilde{\underline{a}}|\underline{a}}[\widetilde{a}_{\ell,k} \cdot x]}{m_0} \right] \leq \frac{1}{m_0}.$$

This shows that $\|f - \mathbb{E}\overline{f}_{m,m_0}\|_2^2 \leq \frac{v^2}{m_0^2}$. Since these bounds hold on average, there exists a realization of (29) for which the bounds are also valid. Note that the vector $\frac{1}{m_0} \sum_{\ell=1}^{m_0} \widetilde{a}_{\ell,k}$ has ℓ^0 norm at most m_0 and unit ℓ^1 norm.

The fact that the bounds also hold for f adjusted by a linear or quadratic term (under an assumption of finite $v_{f,2}$ or $v_{f,3}$) follows from (21) and (23). □

Acknowledgment

The authors would like to thank the anonymous reviewers whose detailed feedback led to dramatic improvements to this paper. They also thank Joowon Kim for helpful comments on earlier drafts of this manuscript.

References

[1] Andrew R. Barron. Neural net approximation. *Yale Workshop on Adaptive and Learning Systems, Yale University Press*, 1992.

[2] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.

[3] Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

[4] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Inform. Theory*, 44(2):525–536, 1998.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[6] Leo Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory*, 39(3):999–1013, 1993.

[7] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. *arXiv Preprint*, February, 2017.

[8] Gerald H. L. Cheang and Andrew R. Barron. A better approximation for balls. *J. Approx. Theory*, 104(2):183–203, 2000.

[9] David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, 69(2):217–232, 1995.

[10] Stratis Ioannidis and Andrea Montanari. Learning combinations of sigmoids through gradient estimation. *arXiv preprint arXiv:1708.06678*, 2017.

[11] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[12] Jason M. Klusowski and Andrew R. Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *Submitted*, 2018.

[13] Jason M. Klusowski and Andrew R. Barron. Minimax lower bounds for ridge combinations including neural nets. *Proceedings IEEE International Symposium on Information Theory*, Aachen, Germany, pages 1377–1380, June, 2017.

[14] Věra Kůrková and Marcello Sanguineti. Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets. *Discrete Appl. Math.*, 155(15):1930–1942, 2007.

[15] Y. Makovoz. Random approximants and neural networks. *J. Approx. Theory*, 85(1):98–109, 1996.

[16] Y. Makovoz. Uniform approximation by neural networks. *J. Approx. Theory*, 95(2):215–228, 1998.

[17] Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

[18] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

[19] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[20] V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971.

[21] Joseph E. Yukich, Maxwell B. Stinchcombe, and Halbert White. Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Trans. Inform. Theory*, 41(4):1021–1027, 1995.

[22] Yuchen Zhang, Jason D Lee, Martin J Wainwright, and Michael I Jordan. Learning halfspaces and neural networks with random initialization. *arXiv preprint arXiv:1511.07948*, 2015.