Uniform Approximation by Neural Networks Activated by First and Second Order Ridge Splines

Jason M. Klusowski, *Student Member, IEEE* and Andrew R. Barron, *Fellow, IEEE*

***Abstract*—We establish sup-norm error bounds for functions that are approximated by linear combinations of first and second order ridge splines and show that these bounds are near-optimal.**

***Index Terms*—Artificial neural networks, approximation error, spline, stratified sampling**

I. Introduction

FUNCTIONS defined on $D = [-1, 1]^d$ are approximated using linear combinations of ridge functions with one layer of nonlinearities. These approximations are employed via functions of the form

$$f_m(x) = f_m(x, \zeta) = \sum_{k=1}^{m} \beta_k \phi(\alpha_k \cdot x + t_k), \tag{I.1}$$

which are parameterized by the vector ζ, consisting of α_k in $\mathbb{R}^d$, and t_k, β_k in $\mathbb{R}$ for $k = 1, \dots, m$, where $m \geqslant 1$ is the number of nonlinear terms. The function ϕ is allowed to be quite general. For example, it can be bounded and Lipschitz, polynomials with certain controls on their degrees, or bounded with jump discontinuities. Recently in [1], it has been shown how ramp functions $\phi(x) = (x)_+ = 0 \vee x$ can be used to give desirable $L^2(D)$ risk bounds of order $(\log d/n)^{1/4}$, useful even when $d \gg n$, or $((d/n)\log(n/d))^{1/2+1/(2(d+1))}$ for estimating a function f, given observations $\{(X_i, Y_i)\}_{i=1}^n$ in a regression setting $Y_i = f(X_i) + \epsilon_i$. These bounds take advantage of the fact that if f satisfies a certain spectral norm condition, then f_m with ϕ a ramp function and $\|\alpha_k\|_1$, $|t_k|$, and $\sum_{k=1}^m |\beta_k|$ bounded serves as a particularly useful approximator of f. In this case, $\phi(\alpha \cdot x + t)$ is Lipschitz with respect to α and t, and the boundedness of $\|\alpha\|_1$ and $|t|$ yields bounded sup-norm covering numbers for their respective norm spaces. Note that such properties are not enjoyed by step functions $\phi(x) = \mathbb{I}\{x > 0\}$ and modeling them using ramp functions requires unbounded internal parameters because $(\tau x)_+ \wedge 1 \to \mathbb{I}\{x > 0\}$ as $\tau \to +\infty$. According to the classic theory [2], [3], if the domain of f is contained in a hyper-cube $[-1, 1]^d$ and f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix\cdot\omega} \tilde{f}(\omega) d\omega$, then the spectral condition $v_{f,1} < \infty$, where $v_{f,s} = \int_{\mathbb{R}^d} \|\omega\|_1^s |\tilde{f}(\omega)| d\omega$, is enough to ensure that f can be approximated in $L^\infty(D)$ by equally weighted ($\beta_1 = \dots = \beta_m$) linear combinations of functions of the form (I.1) with $\phi(x) = \mathbb{I}\{x > 0\}$. Typical rates of an m-term approximation (I.1) are at most $cv_{f,1}\sqrt{d} m^{-1/2}$, where c is a universal constant [2], [4], [5].

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

Unlike the case with step activation functions, our analysis makes no use of the combinatorial properties of half-spaces as in Vapnik-Chervonenkis theory [6], [7] to obtain covering numbers of relevant spaces. The $L^2(D)$ case for ramp ridge functions (also known as hinging hyperplanes) was considered in [8] and our $L^\infty(D)$ bounds improve upon that line of work.

In this paper, we will show that even tighter rates of approximation are possible under two different conditions: $v_{f,2}$ and $v_{f,3}$ finite. Interestingly, there is a disparity in the quality and proof technique of the upper bounds depending on the form of the weights β_k and degree of smoothness of the activation function. The main idea we use for our results originates from [9] and [10] and is essentially stratified sampling with proportional allocation. This technique is widely applied in survey sampling as a means of variance reduction [11].

At the end, we will also discuss the degree to which these bounds can be improved. Throughout this paper, we will state explicitly how our bounds depend on d so that the reader can fully appreciate the complexity of approximation.

II. Statement of results

Theorem 1. *Suppose f admits the integral representation*

$$f(x) = \int_{[0,1]\times\mathbb{S}^{d-1}} s(t, \alpha)\ (\alpha \cdot x - t)_+^q d\mu(t \times \alpha),$$

for x in $D = [-1, 1]^d$, where μ is a sub-stochastic measure on $[0, 1] \times \mathbb{S}^{d-1}$, $s(t, \alpha)$ is either -1 or $+1$, and $q = 1, 2$. There exists a linear combination of ramp ridge functions of the form

$$f_m(x) = \frac{v}{m} \sum_{k=1}^{m} \beta_k (x \cdot \alpha_k - t_k)_+^q \tag{II.1}$$

with $\beta_k \in [-1, 1]$, $\|\alpha_k\|_1 = 1$, $0 \leqslant t_k \leqslant 1$, and $v \leqslant 1$ such that

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

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

$$((\log m)^{1-q/2} \vee \sqrt{d}) m^{-1/2-1/(d+2)}.$$

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

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

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

$$f_m(x) = \beta_0 + x\cdot\alpha_0 + \frac{v}{m}\sum_{k=1}^{m}\beta_k(x\cdot\alpha_k - t_k)_+ \qquad \text{(II.2)}$$

with $\beta_k \in [-1,1]$, $\|\alpha_k\|_1 = 1$, $0 \leqslant t_k \leqslant 1$, $\beta_0 = f(0)$, $\alpha_0 = \nabla f(0)$*, and* $v \leqslant 2v_{f,2}$ *such that*

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

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

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

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

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

There exists a linear combination of second order spline ridge functions of the form

$$f_m(x) = \beta_0 + x\cdot\alpha_0 + x^T A_0 x + \frac{v}{2m}\sum_{k=1}^{m}\beta_k(x\cdot\alpha_k - t_k)_+^2 \qquad \text{(II.3)}$$

with $\beta_k \in [-1,1]$, $\|\alpha_k\|_1 = 1$, $0 \leqslant t_k \leqslant 1$, $\beta_0 = f(0)$, $\alpha_0 = \nabla f(0)$, $A_0 = \nabla\nabla^T f(0)$*, and* $v \leqslant 2v_{f,3}$ *such that*

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

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

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

Let $\mathcal{H}_q = \{x \mapsto z(\alpha\cdot x - t)_+^q : \|\alpha\|_1 = 1,\ 0 \leqslant t \leqslant 1,\ z \in \{-1,+1\}\}$ and for $p \in [2,\infty]$ let $\mathcal{F}_p^q$ denote the closure of the convex hull of $\mathcal{H}_q$ 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 $p = \infty$. By Hölder's inequality, we have the inclusions $\mathcal{F}_\infty^q \subset \cdots \subset \mathcal{F}_2^q$. We let $\mathcal{C}_m$ denote the collection of all combinations $\sum_{h\in\mathcal{H}}\beta_h h$ with positive weights β_h that sum to at most one and at most m that are non-zero. By Theorems 2 and 3, after possibly subtracting a linear or quadratic part, $f/v_{f,q+1}$ belongs to $\mathcal{F}_p^q$. For $p \in [2,\infty]$ and $\epsilon > 0$, we define the ϵ-metric entropy $M_p(\epsilon)$ to be the logarithm of the quantity

$$\min\{n : \exists\tilde{\mathcal{F}} \subset \mathcal{F}_p^q,\ |\tilde{\mathcal{F}}| = n,\ \text{s.t.} \inf_{\tilde{f}\in\tilde{\mathcal{F}}}\sup_{f\in\mathcal{F}_p^q}\|f - \tilde{f}\|_p < \epsilon\}.$$

Theorem 1 implies that

$$\inf_{f_m\in\mathcal{C}_m^q}\sup_{f\in\mathcal{F}_\infty^p}\|f - f_m\|_\infty \leqslant c((\log m)^{1-q/2}\vee\sqrt{d})m^{-1/2-1/d},$$

for some universal constant $c > 0$.

Remark 1. *In [10], it was shown that the standard order* $m^{-1/2}$ *bound alluded to earlier could be improved to be of order* $m^{-1/2-1/(2d)}\sqrt{\log m}$ *under the slightly stronger condition of finite* $v_{f,1}^\star = \sup_{u\in\mathbb{S}^{d-1}}\int_0^\infty r^d|\tilde{f}(ru)|dr$*. In general, our assumptions are not stronger than this since the function* f *with Fourier transform* $\tilde{f}(\omega) = e^{-\|\omega-\omega_0\|}/\|\omega-\omega_0\|$ *for* $\omega_0 \neq 0$ *and* $d \geqslant 2$ *has infinite* $v_{f,1}^\star$ *but finite* $v_{f,s}$ *for* $s \geqslant 0$*. However, the function* f *with Fourier transform* $\tilde{f}(\omega) = 1/(1+\|\omega\|)^{d+2}$ *has finite* $v_{f,1}^\star$ *but infinite* $v_{f,s}$ *for* $s \geqslant 2$*.*

We will only prove Theorems 2 and 3. The proof of Theorem 1 uses similar arguments and so we omit it here.

Proof of Theorem 2: If $|z| \leqslant c$, we note the identity

$$-\int_0^c[(z-u)_+e^{iu} + (-z-u)_+e^{-iu}]du = e^{iz} - iz - 1. \qquad \text{(II.4)}$$

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

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

Multiplying the above by $\tilde{f}(\omega)$, 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

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

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

$$p(z,t,\omega) = |\cos(z\|\omega\|_1 t + b(\omega))|\|\omega\|_1^2|\tilde{f}(\omega)|/v, \qquad \text{(II.5)}$$

where

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

Consider a random variable $h(z,t,\alpha)(x)$ that equals

$$(z\alpha\cdot x - t)_+\ s(zt,\omega),$$

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

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

Considering a maximal ϵ-packing of the space $\Omega = \{(z,s,t,\alpha)' : (z,s)' \in \{-1,+1\}^2,\ 0 \leqslant t \leqslant 1,\ \|\alpha\|_1 = 1\}$ with respect to the norm

$$\|(z,s,t,\alpha)'\| = |z| + |s| + |t| + \|\alpha\|_1$$

consisting of M balls $\mathcal{B}_1,\dots,\mathcal{B}_M$ with radius ϵ. We can show that $4(2/\epsilon)^d \leqslant M \leqslant 4(5/\epsilon)^d$ as follows. The space $\{-1,+1\}^2$ has an ϵ-packing number equal to 4 for $\epsilon < 2$ (with respect to Hamming distance), $\{t : 0 \leqslant t \leqslant 1\}$ has an $\epsilon/2$-packing number between $2/\epsilon$ and $(2/\epsilon+1) \leqslant (3/\epsilon)$, $\epsilon \leqslant 1$ (with respect to $|\cdot|$), and $\{\alpha : \|\alpha\|_1 = 1\}$ has an $\epsilon/2$-packing number between $(2/\epsilon)^d$ and $(4/\epsilon+1)^{d-1} \leqslant (5/\epsilon)^{d-1}$, $\epsilon \leqslant 1$ (with respect to

$\|\cdot\|_1$). Thus Ω has an ϵ-packing number between $4(2/\epsilon)^d$ and $4(3/\epsilon)(5/\epsilon)^{d-1} \leqslant 4(5/\epsilon)^d$. For $k = 1, \ldots, M$ define

$$p_k(z,t,\omega) = p(z,t,\omega)\mathbb{I}\{(z, s(zt,\omega), t, \alpha(\omega))' \in \mathcal{B}_k\}/L_k,$$

where L_k is chosen to make p_k integrate to one. A very important property we will use is that if $(z,t,\omega)'$ is distributed according to p_k and if $(z_k, t_k, \omega_k)'$ denotes the center of $\mathcal{B}_k$, then

$$\begin{aligned} |h(z,t,\alpha)(x) - h(z_k,t_k,\alpha_k)(x)| &\leqslant \|\alpha - \alpha_k\|_1 \|x\|_\infty + |t - t_k| \\ &\leqslant \epsilon. \end{aligned} \tag{II.6}$$

Define a sequence of M independent random variables $\{m_k\}_{1 \leqslant k \leqslant M}$ 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, $\underline{m} = \{m_k\}_{1\leqslant k \leqslant M}$, take a random sample $\underline{\gamma} = \{(z_{i,k}, t_{i,k}, \alpha(\omega_{i,k}))'\}_{1 \leqslant i \leqslant n_k,\, 1 \leqslant k \leqslant M}$ of size $n_k = m_k + \mathbb{I}\{m_k = 0\}$ 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 ($m_1, \ldots, m_M$). The within-stratum variability is now smaller than the population variability as evidenced by (II.6).

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

$$\begin{aligned} \sum_{k=1}^M n_k &= \sum_{k=1}^M m_k \mathbb{I}\{m_k > 0\} + \sum_{k=1}^M \mathbb{I}\{m_k = 0\} \\ &\leqslant \sum_{k=1}^M (mL_k + 1)\mathbb{I}\{m_k > 0\} + \sum_{k=1}^M \mathbb{I}\{m_k = 0\} \\ &= m \sum_{k=1}^M L_k \mathbb{I}\{m_k > 0\} + M \\ &\leqslant m + M. \end{aligned} \tag{II.7}$$

For $i = 1, \ldots, m_k$, let $h_{i,k} = h(z_{i,k}, t_{i,k}, \alpha(\omega_{i,k}))$ and $f_k = \frac{2vm_k}{mn_k} \sum_{i=1}^{n_k} h_{i,k}$. Also, let $\overline{f} = \sum_{k=1}^M f_k$. A simple calculation shows that the mean of $\overline{f}$ is $f - f(0) - x \cdot \nabla f(0)$, which we denote by g. Next, we upper bound

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

by

$$\begin{aligned} &\frac{v}{m} \mathbb{E}_{\underline{m}} \sup_{x \in D} |\sum_{k=1}^M (m_k - L_k m) \mathbb{E}_{p_k} h(x)| + \\ &\frac{v}{m} \mathbb{E}_{\underline{m}} \mathbb{E}_{\underline{\gamma}|\underline{m}} \sup_{x \in D} |\sum_{k=1}^M \sum_{i=1}^{n_k} \frac{m_k}{n_k} (h_{i,k}(x) - \mathbb{E}_{p_k} h(x))| \end{aligned} \tag{II.8}$$

Now

$$\begin{aligned} &\mathbb{E}_{\underline{\gamma}|\underline{m}} \sup_{x \in D} |\sum_{k=1}^M \sum_{i=1}^{n_k} \frac{m_k}{n_k} (h_{i,k}(x) - \mathbb{E}_{p_k} h(x))| \leqslant \\ &2\mathbb{E}_{\underline{\gamma}|\underline{m}} \sup_{x \in D} |\sum_{k=1}^M \sum_{i=1}^{n_k} \sigma_{i,k} \frac{m_k}{n_k} [h_{i,k}(x) - \mu_{i,k}(x)]|, \end{aligned} \tag{II.9}$$

where $\{\sigma_{i,k}\}$ is a sequence of independent identically distributed Rademacher variables and $\{x \mapsto \mu_{i,k}(x)\}$ is any sequence of functions defined on D [see for example Lemma 2.3.6 in [12]]. For notational brevity, we define $\tilde{h}_{i,k}(x) = \frac{m_k}{n_k}[h_{i,k}(x) - \mu_{i,k}(x)]$. By Dudley's entropy integral method [see Corollary 13.2 in [13]], the quantity in (II.9) can be bounded by

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

where $N(u,D)$ is the u-metric entropy of D with respect to the norm $d(x,x')$ defined by

$$\begin{aligned} d^2(x,x') &= \sum_{k=1}^M \sum_{i=1}^{n_k} (\tilde{h}_{i,k}(x) - \tilde{h}_{i,k}(x'))^2 \\ &\leqslant (m+M)\|x - x'\|_\infty^2, \end{aligned} \tag{II.11}$$

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

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

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

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

where $\{\sigma_k\}$ is a sequence of independent identically distributed Rademacher variables. For notational brevity, we write $\tilde{h}_k(x) = (m_k - L_k m)\mathbb{E}_{p_k} h(x)$. We can also bound (II.13) by (II.10), except this time $N(u,D)$ is the u-metric entropy of D with respect to the norm $\rho^2(x,x')$ defined by

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

where the last line follows from $|m_k - L_k m| \leqslant 1$ and $|\mathbb{E}_{p_k} h(x) - \mathbb{E}_{p_k} h(x')| \leqslant \|x - x'\|_\infty$. The quantity δ is also less than $2\sqrt{M}$, since $\sup_{x \in D} |\tilde{h}_k(x)| \leqslant 2$ and moreover $N(u,D) \leqslant d\log(3\sqrt{M}/u)$. Evaluating the integral in (II.10) with these specifications yields a bound on the first term in (II.8) of

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

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

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

Choose

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

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

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

We showed earlier that $M \asymp \epsilon^{-d}$. Thus the stipulation $M \leqslant m$ implies that ϵ is of order $m^{-1/(d+2)}$. Since the inequality (II) holds on average, there is a realization of $\overline{f}$ for which $\sup_{x\in D}|\overline{f}(x)-g(x)|$ has the same bound. Note that $\overline{f}$ 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 (II.8) is zero and hence $\mathbb{E}\sup_{x\in D}|\overline{f}(x)-g(x)|$ is not greater than (II.12). The conclusion follows with $M = m$ and ϵ of order $m^{-1/d}$. ■

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

$$\int_{\mathbb{R}^d}(e^{i\omega\cdot x}+(\omega\cdot x)^2/2-i\omega\cdot x-1)\tilde{f}(\omega)d\omega. \qquad \text{(II.16)}$$

As before, the integrand in (II.16) admits an integral representation given by

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

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

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

where

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

and

$$p(z,t,\omega)=|\sin(z\|\omega\|_1 t+b(\omega))|\|\omega\|_1^3|\tilde{f}(\omega)|/v,$$

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

The metric $d(x,x')$ is in fact bounded by a constant multiple of $\sqrt{m+M}\epsilon\|x-x'\|_\infty$. To see this, we note that the function $\tilde{h}_{i,k}(x)$ has the form (up to a sign difference)

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

with $\|\alpha-\alpha_k\|_1+|t-t_k|<\epsilon$. The gradient of $\tilde{h}_{i,k}(x)$ with respect to x is equal to

$$\nabla\tilde{h}_{i,k}(x)=\frac{2m_k}{n_k}[(\alpha(\alpha\cdot x-t)_+-\alpha_k(\alpha_k\cdot x-t_k)_+].$$

Adding and subtracting $\frac{2m_k}{n_k}\alpha(\alpha_k\cdot x-t_k)_+$ to the above expression yields the bound of order ϵ for $\sup_{x\in D}\|\nabla\tilde{h}_{i,k}(x)\|_1$. Taylor's theorem yields the desired bound on $d(x,x')$. Again using Dudley's entropy integral, we can bound $\mathbb{E}\sup_{x\in D}|\overline{f}(x)-g(x)|$ 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 results follow with $M = m\epsilon^2/(1-\epsilon^2)\leqslant m$ and ϵ of order $m^{-1/(d+2)}$ or $M = m$ and ϵ of order $m^{-1/d}$. 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. Note that this rate is the same as what was achieved in Theorem 2, without a $\sqrt{(\log m)/d}$ factor. ■

Next, we investigate the optimality of the above rates.

Theorem 4. *For* $p\in[2,\infty]$,

$$\inf_{f_m\in\mathcal{C}_m^q}\sup_{f\in\mathcal{F}_p^q}\|f-f_m\|_p\geqslant (c(d^{2q+3}\log d)(m\log m))^{-1/2-(q+1)/d},$$

for some universal positive constant c.

Ignoring the dependence on d and logarithmic factors in m, this result coupled with Theorem 1 implies that $\inf_{f_m\in\mathcal{C}_m^1}\sup_{f\in\mathcal{F}_p^1}\|f-f_m\|_p$ is between $m^{-1/2-2/d}$ and $m^{-1/2-1/d}$.

We now use a result that is contained in Lemma 4.2 in [14].

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

(i) *card*$(G):=|G|\geqslant 3$,

(ii) $\sum_{g,g'\in G,\ g\neq g'}|\langle g,g'\rangle|\leqslant\delta^2$

(iii) $\delta^2\leqslant\|G\|^2:=\min_{g\in G}\|g\|^2$

Then there exists a collection $\Omega\subset\{0,1\}^{|G|}$ with cardinality at least $2^{(1-H(1/4))|G|-1}$, 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 $\mathcal{F}_G=\left\{\frac{1}{|G|}\sum_{g\in G}\beta_g g:\beta\in\Omega\right\}$ is separated by at least $\frac{1}{2}\sqrt{\frac{\|G\|^2-\delta^2}{|G|}}$ in $\|\cdot\|$.

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

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

satisfies the assumption of the previous lemma with $H=L^2(D,P)$, where P is the uniform probability measure on D. Moreover, $|G|=r^d$, $\|G\|=1/(4\sqrt{2}\pi d^2r^2)$, and $\mathcal{F}_G\subset\mathcal{F}_p^1$ for all $p\in[2,\infty]$.

Proof: We first observe the identity

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

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

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

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 G 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 $\sin(\pi\theta \cdot x) = \frac{1}{2i}(\prod_{k=1}^d e^{i\pi\theta_k x_k} - \prod_{k=1}^d e^{-i\pi\theta_k x_k})$. ■

An important consequence of using ramp activation functions that is not available if one uses step activation functions is the ability to bound the metric entropy $M_\infty(\epsilon)$ of the approximation classes.

The Lipschitz property of the ramp ridge function with respect to its internal parameters implies that if

$$\inf_{f_m \in \mathcal{C}_m^1} \sup_{f \in \mathcal{F}_p^1} \|f - f_m\|_p < r(m) = \epsilon/2,$$

then there exists a cover $\tilde{\mathcal{C}}_m^1$ of $\mathcal{C}_m^1$ with cardinality at most $(9/\epsilon)^m \binom{(9/\epsilon)^d + m - 1}{m}$ such that

$$\inf_{f_m \in \tilde{\mathcal{C}}_m^1} \sup_{f \in \mathcal{F}_p^1} \|f - f_m\|_p < \epsilon.$$

This shows that

$$M_p(\epsilon) \leqslant c_0 dm \log(1/\epsilon),$$

for some positive universal constant c_0.

Proof of Theorem 4: We only give the proof for $q = 1$. The other case is handled similarly. Suppose contrary to the hypothesis,

$$\inf_{f_m \in \mathcal{C}_m^1} \sup_{f \in \mathcal{F}_p^1} \|f - f_m\|_p < (c(d^5 \log d)(m \log m))^{-1/2-2/d} = \epsilon/2,$$

for some universal constant $c > 0$ to be chosen later. By the previous argument, $M_p(\epsilon) \leqslant c_1 dm \log(cdm)$ for some positive universal constant $c_1 > 0$. However, using Lemma 2 with

$$\frac{1}{2}\frac{\|G\|}{\sqrt{|G|}} = \frac{1}{8\sqrt{2}\pi d^2 r^{2+d/2}} = \epsilon/2$$

determines $r = (c_2 d^2 \epsilon)^{-2/(d+4)}$, for some universal constant $c_2 > 0$. Thus a valid lower bound for $M_2(\epsilon)$ is $|G| = r^d \geqslant c_3 cdm \log d \log m$ for some universal constant $c_3 > 0$. Since $M_p(\epsilon) \geqslant M_2(\epsilon)$ and $\mathcal{F}_G \subset \mathcal{F}_p^1$, we have

$$c_1 dm \log(cdm) \geqslant c_3 cdm \log d \log m.$$

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

ACKNOWLEDGMENT

The authors would like to thank Joowon Kim for helpful comments on earlier drafts of this manuscript.

REFERENCES

[1] J. M. Klusowski and A. R. Barron, "Risk bounds for high-dimensional ridge function combinations including neural networks," *Submitted to Annals of Statistics*, 2016.

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

[3] ——, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 930–945, 1993. [Online]. Available: http://dx.doi.org/10.1109/18.256500

[4] G. H. L. Cheang and A. R. Barron, "A better approximation for balls," *J. Approx. Theory*, vol. 104, no. 2, pp. 183–203, 2000. [Online]. Available: http://dx.doi.org/10.1006/jath.1999.3441

[5] J. E. Yukich, M. B. Stinchcombe, and H. White, "Sup-norm approximation bounds for networks through probabilistic methods," *IEEE Trans. Inform. Theory*, vol. 41, no. 4, pp. 1021–1027, 1995. [Online]. Available: http://dx.doi.org/10.1109/18.391247

[6] V. N. Vapnik and A. J. Červonenkis, "The uniform convergence of frequencies of the appearance of events to their probabilities," *Teor. Verojatnost. i Primenen.*, vol. 16, pp. 264–279, 1971.

[7] D. Haussler, "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension," *J. Combin. Theory Ser. A*, vol. 69, no. 2, pp. 217–232, 1995. [Online]. Available: http://dx.doi.org/10.1016/0097-3165(95)90052-7

[8] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 999–1013, 1993. [Online]. Available: http://dx.doi.org/10.1109/18.256506

[9] Y. Makovoz, "Random approximants and neural networks," *J. Approx. Theory*, vol. 85, no. 1, pp. 98–109, 1996. [Online]. Available: http://dx.doi.org/10.1006/jath.1996.0031

[10] ——, "Uniform approximation by neural networks," *J. Approx. Theory*, vol. 95, no. 2, pp. 215–228, 1998. [Online]. Available: http://dx.doi.org/10.1006/jath.1997.3217

[11] J. 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*, vol. 97, no. 4, pp. 558–625, 1934.

[12] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes*, ser. Springer Series in Statistics. Springer-Verlag, New York, 1996, with applications to statistics. [Online]. Available: http://dx.doi.org/10.1007/978-1-4757-2545-2

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities*. Oxford University Press, Oxford, 2013, a nonasymptotic theory of independence, With a foreword by Michel Ledoux. [Online]. Available: http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001

[14] V. Kůrková and M. Sanguineti, "Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets," *Discrete Appl. Math.*, vol. 155, no. 15, pp. 1930–1942, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.dam.2007.04.007