Minimax Lower Bounds for Ridge Combinations Including Neural Nets

Jason M. Klusowski and Andrew R. Barron

Department of Statistics

Yale University

New Haven, CT, USA

Email: {jason.klusowski, andrew.barron}@yale.edu

Abstract—**Estimation of functions of d variables is considered using ridge combinations of the form $\sum_{k=1}^{m} c_{1,k}\phi(\sum_{j=1}^{d} c_{0,j,k}x_j - b_k)$ where the activation function ϕ is a function with bounded value and derivative. These include single-hidden layer neural networks, polynomials, and sinusoidal models. From a sample of size n of possibly noisy values at random sites $X \in B = [-1, 1]^d$, the minimax mean square error is examined for functions in the closure of the ℓ_1 hull of ridge functions with activation ϕ. It is shown to be of order d/n to a fractional power (when d is of smaller order than n), and to be of order $(\log d)/n$ to a fractional power (when d is of larger order than n). Dependence on constraints v_0 and v_1 on the ℓ_1 norms of inner parameter c_0 and outer parameter c_1, respectively, is also examined. Also, lower and upper bounds on the fractional power are given. The heart of the analysis is development of information-theoretic packing numbers for these classes of functions.**

Index Terms—**Nonparametric regression; nonlinear regression; neural nets; penalization; machine learning; high-dimensional data analysis; learning theory; generalization error; greedy algorithms; metric entropy; packing sets; polynomial nets; sinusoidal nets; constant weight codes**

I. INTRODUCTION

Ridge combinations provide flexible classes for fitting functions of many variables. The ridge activation function may be a general Lipschitz function. When the ridge activation function is a sigmoid, these are single-hidden layer artificial neural nets. When the activation is a sine or cosine function, it is a sinusoidal model in a ridge combination form. We consider also a class of polynomial nets which are combinations of Hermite ridge functions. Ridge combinations are also the functions used in projection pursuit regression fitting. What distinguishes these models from other classical functional forms is the presence of parameters internal to the ridge functions which are free to be adjusted in the fit. In essence, it is a parameterized, infinite dictionary of functions from which we make linear combinations. This provides a flexibility of function modeling not present in the case of a fixed dictionary. Here we discuss results on risk properties of estimation of functions using these models and we develop new minimax lower bounds.

For a given activation function $\phi(z)$ on $\mathbb{R}$, consider the parameterized family $\mathcal{F}_m$ of functions

$$f_m(x) = f_m(x, c_0, c_1, b) = \sum_{k=1}^{m} c_{1,k}\phi(\sum_{j=1}^{d} c_{0,j,k}x_j - b_k), \tag{1}$$

where $c_1 = (c_{1,1}, \ldots, c_{1,m})'$ is the vector of outer layer parameters and $c_{0,k} = (c_{0,1,k}, \ldots, c_{0,d,k})'$ are the vectors of inner parameters for the single hidden-layer of functions $\phi(c_{0,k} \cdot x - b_k)$ with horizontal shifts $b = (b_1, \ldots, b_m)$, $k = 1, \ldots, m$. For positive v_0, let

$$\mathcal{D}_{v_0} = \mathcal{D}_{v_0,\phi} = \{\phi(\theta \cdot x - t), \ x \in B : \|\theta\|_1 \leqslant v_0, \ t \in \mathbb{R}\} \tag{2}$$

be the dictionary of all such inner layer ridge functions $\phi(\theta \cdot x - t)$ with parameter restricted to the ℓ_1 ball of size v_0 and variables x restricted to the cube $[-1, 1]^d$. The choice of the ℓ_1 norm on the inner parameters is natural as it corresponds to $\|\theta\|_B = \sup_{x \in B} |\theta \cdot x|$ for $B = [-1, 1]^d$.

Let $\mathcal{F}_{v_0,v_1} = \mathcal{F}_{v_0,v_1,\phi} = \ell_1(v_1, \mathcal{D}_{v_0})$ be the closure of the set of all linear combinations of functions in $\mathcal{D}_{v_0}$ with ℓ_1 norm of outer coefficients not more than v_1. These v_0 and v_1 control the freedom in the size of this function class. They can either be fixed for minimax evaluations, or adapted in the estimation (as reflected in some of the upper bounds on risk for penalized least square estimation). The functions of the form (1) are in $\ell_1(v_1, \mathcal{D})$ when $\|c_{0,k}\|_1 \leqslant v_0$ and $\|c_1\|_1 \leqslant v_1$. Indeed, let $\mathcal{F}_{m,v_0,v_1} = \ell_1(m, v_1, \mathcal{D}_{v_0})$ be the subset of such functions in $\ell_1(v_1, \mathcal{D}_{v_0})$ that use m terms.

Data are of the form $\{(X_i, Y_i)\}_{i=1}^{n}$, drawn independently from a joint distribution $P_{X,Y}$ with P_X on $[-1, 1]^d$. The target function is $f(x) = \mathbb{E}[Y|X = x]$, the mean of the conditional distribution $P_{Y|X=x}$, optimal in mean square for the prediction of future Y from corresponding input X. In some cases, assumptions are made on the error of the target function $\epsilon_i = Y_i - f(X_i)$ (i.e. bounded, Gaussian, or sub-Gaussian).

From the data, estimators $\hat{f}(x) = \hat{f}(x, \{(X_i, Y_i)\}_{i=1}^{n})$ are formed and the loss at a target f is the $L_2(P_X)$ square error $\|f - \hat{f}\|^2$ and the risk is the expected squared error $\mathbb{E}\|f - \hat{f}\|^2$. For any class of functions $\mathcal{F}$ on $[-1, 1]^d$, the minimax risk is

$$R_{n,d}(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\|f - \hat{f}\|^2, \tag{3}$$

where the infimum runs over all estimators $\hat{f}$ of f based on the data $\{(X_i, Y_i)\}_{i=1}^n$.

It is known that for certain complexity penalized least squares estimators [1], [2], [3], [4] the risk satisfies

$$\mathbb{E}\|f - \hat{f}\|^2 \leqslant \inf_{f_m \in \mathcal{F}_m} \{\|f - f_m\|^2 + \tfrac{cmd\log n}{n}\}, \qquad (4)$$

where the constant c depends on parameters of the noise distribution and on properties of the activation function ϕ, which can be a step function or a fixed bounded Lipschitz function. The $d\log n$ in the second term is from the log-cardinality of customary d-dimensional covers of the dictionary. The right side is an index of resolvability expressing the tradeoff between approximation error $\|f - f_m\|^2$ and descriptive complexity $md\log n$ relative to sample size, in accordance with risk bounds for minimum description length criteria [5], [6], [7], [8]. When the target f is in $\mathcal{F}_{v_1, v_0}$, it is known as in [9], [10], [11] that $\|f - f_m\|^2 \leqslant v_1^2/m$ with slight improvements possible depending on the dimension $\|f - f_m\|^2 \leqslant v_1^2/m^{1/2+1/d}$ as in [12], [13], [14]. When f is not in $\mathcal{F}_{v_0, v_1}$, let f_{v_0, v_1} be its projection onto this convex set of functions. Then the additional error beyond $\|f - f_{v_0, v_1}\|^2$ is controlled by the bound [1]

$$\inf_m \{\tfrac{v_1^2}{m} + \tfrac{c_1 md \log n}{n}\} = 2v_1\left(\tfrac{c_1 d \log n}{n}\right)^{1/2}. \qquad (5)$$

Moreover, with $\hat{f}$ restricted to $\mathcal{F}_{v_0, v_1}$, this bounds the mean squared error $\mathbb{E}\|\hat{f} - f_{v_0, v_1}\|^2$ from the projection. The same risk is available from ℓ_1 penalized least square estimation [3], [6], [7], [13] and from greedy implementations of complexity and ℓ_1 penalized estimation [3], [13]. The slight approximation improvements (albeit not known whether available by greedy algorithms) provide the risk bound [13]

$$R_{n,d}(\mathcal{F}_{v_0, v_1}) \leqslant c_2\left(\tfrac{dv_0^2 v_1^2}{n}\right)^{1/2+1/(2(d+1))}, \qquad (6)$$

for bounded Lipschitz activation functions ϕ, improving a similar result in [15], [14]. This fact can be shown through improved upper bounds on the metric entropy from [16].

A couple of lower bounds on the minimax risk in $\mathcal{F}_{v_0, v_1}$ are known [14] and, improving on [14], the working paper [13] states the lower bound

$$R_{n,d}(\mathcal{F}_{v_0, v_1}) \geqslant c_3 v_1^{d/(d+2)}\left(\tfrac{1}{d^4 n}\right)^{1/2+1/(d+2)} \qquad (7)$$

for an unconstrained v_0.

Note that for large d, these exponents are near $1/2$. Indeed, if d is large compared to $\log n$, then the bounds in (6) and (7) are of the same order as with exponent $1/2$. It is desirable to have improved lower bounds which take the form d/n to a fractional power as long as d is of smaller order than n.

Good empirical performance of neural net (and neural net like) models has been reported as in [17] even when d is much larger than n, though theoretical understanding has been lacking. Current developments [13] obtain upper bounds on risk of the form

$$R_{n,d}(\mathcal{F}_{v_0, v_1}) \leqslant c_4\left(\tfrac{v_0^2 v_1^4 \log(d+1)}{n}\right)^\gamma, \qquad (8)$$

for fixed positive γ, again for bounded Lipschitz ϕ. These allow d much larger than n, as long as $d = e^{o(n)}$. With greedy implementations of least squares over a discretization of the parameter with complexity or ℓ_1 penalty, such upper bounds are obtained in [13] with $\gamma = 1/3$. At the expense of a slightly worse exponent on v_1 and an additional smoothness assumption on ϕ, the rate with $\gamma = 1/3$ is also possible when the greedy algorithm selects candidate neurons from a continuum of choices.

It is desirable likewise to have lower bounds on the minimax risk for this setting that show that is depends primarily on $v_0^\alpha v_1^{2\alpha}/n$ to some power (within $\log d$ factors). It is the purpose of this paper to obtain such lower bounds. Here with $\gamma = 1/2$. Thereby, this paper on lower bounds is to provide a companion to (refinement of) the working paper on upper bounds [13]. Lower bounding minimax risk in non-parametric regression is primarily an information-theoretic problem. This was first observed by [18] and then [19], [20] who adapted Fano's inequality in this setting. Furthermore, [14] showed conditions such that the minimax risk ϵ_n^2 is characterized (to within a constant factor) by solving for the approximation error ϵ^2 that matches the metric entropy relative to the sample size $(\log N(\epsilon))/n$, where $N(\epsilon)$ is the size of the largest ϵ-packing set. Accordingly, the core of our analysis is providing packing sets for $\mathcal{F}_{v_0, v_1}$ for specific choices of ϕ.

II. RESULTS FOR SINUSOIDAL NETS

We now state our main result. In this section, it is for the sinusoidal activation function $\phi(z) = \sqrt{2}\sin(\pi z)$. We consider two regimes: when d is larger than v_0 and visa-versa. In each case, this entails putting a non-restrictive technical condition on either quantity. For d larger than v_0, this condition is

$$\tfrac{d}{v_0} + 1 > \left(c_4 \tfrac{v_1^2 n}{v_0 \log(1+d/v_0)}\right)^{1/v_0}, \qquad (9)$$

and when v_0 is larger than d,

$$\tfrac{v_0}{d} + 1 > \left(c_5 \tfrac{v_1^2 n}{d \log(1+v_0/d)}\right)^{1/d}, \qquad (10)$$

for some positive constants c_4, c_5. Note that when d is large compared to $\log n$, condition (10) holds. Indeed, the left side is at least 2 and the right side is $e^{\frac{1}{d}\log\left(\frac{v_1^2 n}{d\log(1+v_0/d)}\right)}$, which is near 1. Likewise, (9) holds when v_0 is large compared to $\log n$.

Theorem 1. *Consider the model $Y = f(X) + \varepsilon$ for $f \in \mathcal{F}_{v_0, v_1, sine}$, where $\varepsilon \sim N(0, 1)$ and $X \sim Uniform[-1, 1]^d$. If d is large enough so that (9) is satisfied, then*

$$R_{n,d}(\mathcal{F}_{v_0, v_1, sine}) \geqslant c_6\left(\tfrac{v_0 v_1^2 \log(1+d/v_0)}{n}\right)^{1/2}, \qquad (11)$$

for some universal constant $c_6 > 0$. Furthermore, if v_0 is large enough so that (10) is satisfied, then

$$R_{n,d}(\mathcal{F}_{v_0, v_1, sine}) \geqslant c_7\left(\tfrac{dv_1^2 \log(1+v_0/d)}{n}\right)^{1/2}. \qquad (12)$$

for some universal constant $c_7 > 0$.

Before we prove Theorem 1, we first state a lemma which is contained in the proof of Theorem 1 (pp. 46-47) in [21].

Lemma 1. *For integers M, L with $M \geqslant 10$ and $1 \leqslant L \leqslant M/10$, define the set*

$$\mathcal{S} = \{\omega \in \{0,1\}^M : \|\omega\|_1 = L\}.$$

There exists a subset $\mathcal{A} \subset \mathcal{S}$ with cardinality at least $\sqrt{\binom{M}{L}}$ such that the Hamming distance between any pairs of $\mathcal{A}$ is at least $L/5$.

Note that the elements of the set $\mathcal{A}$ in Lemma 1 can be interpreted as binary codes of length M, constant Hamming weight L, and minimum Hamming distance $L/5$. These are called constant weight codes and the cardinality of the largest such codebook, denoted by $A(M, L/5, L)$, is also given a combinatorial lower bound in [22]. The conclusion of Lemma 1 is $A(M, L/5, L) \geqslant \sqrt{\binom{M}{L}}$.

Proof of Theorem 1. For simplicity, we henceforth write $\mathcal{F}_{v_0,v_1}$ instead of $\mathcal{F}_{v_0,v_1,\text{sine}}$. Define the collection $\Lambda = \{\theta \in \mathbb{Z}^d : \|\theta\|_1 \leqslant v_0\}$. Without loss of generality, assume that v_0 is an integer so that $M := \#\Lambda \geqslant \binom{d+v_0}{d}$. Consider sinusoidal ridge functions $\sqrt{2}\sin(\pi\theta \cdot x)$ with θ in Λ. Note that these functions (for $\theta \neq 0$) are orthonormal with respect to the uniform probability measure P on $B = [-1,1]^d$. This fact is easily established using an instance of Euler's formula $\sin(\pi\theta \cdot x) = \frac{1}{2i}(\prod_{k=1}^d e^{i\pi\theta_k x_k} - \prod_{k=1}^d e^{-i\pi\theta_k x_k})$.

For an enumeration $\theta_1, \ldots, \theta_M$ of Λ, define a subclass of $\mathcal{F}_{v_0,v_1}$ by

$$\mathcal{F}_0 = \{f_\omega = \frac{v_1}{L}\sum_{k=1}^M \omega_k \sqrt{2}\sin(\pi\theta_k \cdot x) : \omega \in \mathcal{A}\},$$

where $\mathcal{A}$ is the set in Lemma 1. Any distinct pairs $f_\omega, f_{\omega'}$ in $\mathcal{F}_0$ have $L_2(P)$ squared distance at least $\|f_\omega - f_{\omega'}\|^2 \geqslant v_1^2\|\omega - \omega'\|_2^2/L^2 \geqslant v_1^2/(5L)$. A separation of ϵ^2 determines $L = (v_1/(\sqrt{5}\epsilon))^2$. Depending on the size of d relative to v_0, there are two different behaviors of M. For $d > v_0$, we use $M \geqslant \binom{d+v_0}{v_0} \geqslant (1 + d/v_0)^{v_0}$ and for $d < v_0$, $M \geqslant \binom{d+v_0}{d} \geqslant (1 + v_0/d)^d$.

By Lemma 1, a lower bound on the cardinality of $\mathcal{A}$ is $\sqrt{\binom{M}{L}}$ with logarithm lower bounded by $(L/2)\log(M/L)$. To obtain a cleaner form that highlights the dependence on L, we assume that $L \leqslant \sqrt{M}$, giving $\log(\#\mathcal{A}) \geqslant (L/4)\log M$. Since L is proportional to $(v_1/\epsilon)^2$, this condition puts a lower bound on ϵ of order $v_1 M^{-1/4}$. If $\epsilon > v_1/(1 + d/v_0)^{v_0/4}$, it follows that a lower bound on the logarithm of the packing number is of order $\log N_{d>v_0}(\epsilon) = v_0(v_1/\epsilon)^2 \log(1 + d/v_0)$. If $\epsilon > v_1/(1 + v_0/d)^{d/4}$, a lower bound on the logarithm of the packing number is of order $\log N_{v_0>d}(\epsilon) = d(v_1/\epsilon)^2 \log(1 + v_0/d)$. Thus we have found an ϵ-packing set of these cardinalities. As such, they are lower bounds on the metric entropy of $\mathcal{F}_{v_0,v_1}$.

Next we use the information-theoretic lower bound techniques in [14] or [23]. Let $p_\omega(x,y) = p(x)\psi(y - f_\omega(x))$, where p is

the uniform density on $[-1,1]^d$ and ψ is the $N(0,1)$ density. Then

$$R_{n,d}(\mathcal{F}_{v_0,v_1}) \geqslant (\epsilon^2/4)\inf_{\hat{f}}\sup_{f \in \mathcal{F}_0} \mathbb{P}(\|f - \hat{f}\|^2 \geqslant \epsilon^2),$$

where the estimators $\hat{f}$ are now restricted to $\mathcal{F}_0$. The supremum is at least the uniformly weighted average over $f \in \mathcal{F}_0$. Thus a lower bound on the minimax risk is a constant times ϵ^2 provided the minimax probability is bounded away from zero, as it is for sufficient size packing sets. Indeed, by Fano's inequality as in [14], this minimax probability is at least

$$1 - \frac{\alpha\log(\#\mathcal{F}_0) + \log 2}{\log(\#\mathcal{F}_0)},$$

for α in $(0,1)$, or by an inequality of Pinsker, as in Theorem 2.5 in [23], it is at least

$$\frac{\sqrt{\#\mathcal{F}_0}}{1 + \sqrt{\#\mathcal{F}_0}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log(\#\mathcal{F}_0)}}\right),$$

for some α in $(0, 1/8)$. These inequalities hold provided we have the following

$$\frac{1}{\#\mathcal{F}_0}\sum_{\omega \in \mathcal{A}} D(p_\omega^n \|q) \leqslant \alpha\log(\#\mathcal{F}_0),$$

bounding the mutual information between ω and the data $\{(X_i, Y_i)\}_{i=1}^n$, where q is any fixed joint density for $\{(X_i, Y_i)\}_{i=1}^n$. When suitable metric entropy upper bounds on the log-cardinality of covers $\mathcal{F}_{\omega' \in \mathcal{A}'} := \{f : \|f - f_{\omega'}\| < \epsilon'\}$ of $\mathcal{F}_0$ are available, one may use q as a uniform mixture of $p_{\omega'}^n$ for ω' in $\mathcal{A}'$ as in [14], as long as ϵ and ϵ' are arranged to be of the same order. In the special case that $\mathcal{F}_0$ has small radius already of order ϵ, one has the simplicity of taking $\mathcal{A}'$ to be the singleton set consisting of $\omega' = 0$. In the present case, since each element in $\mathcal{F}_0$ has squared norm $v_1^2/L = 5\epsilon^2$ and pairs of elements in $\mathcal{F}_0$ have squared separation ϵ^2, these function are near $f_0 \equiv 0$ and hence we choose $q = p_0^n$. A standard calculation yields

$$D(p_\omega^n \| p_0^n) \leqslant \frac{n}{2}\|f_\omega\|^2 \leqslant \frac{nv_1^2}{2L} = (5/2)n\epsilon^2.$$

We choose ϵ_n such that this $(5/2)n\epsilon_n^2 \leqslant \alpha\log(\#\mathcal{F}_0)$. Thus, in accordance with [14], if $N_{d>v_0}(\epsilon_n)$ and $N_{v_0>d}(\epsilon_n)$ are available lower bounds on $\#\mathcal{F}_0$, to within a constant factor, a minimax lower bound ϵ_n^2 on the $L_2(P)$ squared error risk is determined by matching

$$\epsilon_n^2 = \frac{\log N_{d>v_0}(\epsilon_n)}{n},$$

and

$$\epsilon_n^2 = \frac{\log N_{v_0>d}(\epsilon_n)}{n}.$$

Solving in either case, we find that

$$\epsilon_n^2 = \left(\frac{v_0 v_1^2 \log(1 + d/v_0)}{n}\right)^{1/2},$$

and

$$\epsilon_n^2 = \left(\frac{dv_1^2 \log(1 + v_0/d)}{n}\right)^{1/2}.$$

These quantities are valid lower bounds on $R_{n,d}(\mathcal{F}_{v_0,v_1})$ to within constant factors, provided $N_{d>v_0}(\epsilon_n)$ and $N_{v_0>d}(\epsilon_n)$ are valid lower bounds on the ϵ_n-packing number of $\mathcal{F}_{v_0,v_1}$.

Checking that $\epsilon_n > v_1/(1 + d/v_0)^{v_0/2}$ and $\epsilon_n > v_1/(1 + v_0/d)^{d/2}$ yields conditions (9) and (10), respectively. $\qquad\square$

Remark. Conditions (9) and (10) are needed to ensure that the lower bounds for the packing numbers take on the form $L \log M$ instead of $L \log(M/L)$. We accomplish this by imposing $L \leqslant \sqrt{M}$. Alternatively, any upper bound of the form M^ρ, $\rho \in (0, 1)$ will work with similar conclusion, adjusting lower bounds (11) and (12) by a factor of $\sqrt{1 - \rho}$, with corresponding adjustment to the requirements on d/v_0 in (9) and v_0/d in (10).

III. IMPLICATIONS FOR NEURAL NETS

The variation of a function f with respect to a dictionary $\mathcal{D}$ [24], also called the atomic norm of f with respect to $\mathcal{D}$, denoted $V_f(\mathcal{D})$, is defined as the infimum of all v such that f is in $\ell_1(v, \mathcal{D})$. Here the closure in the definition of $\ell_1(v, \mathcal{D})$ is taken in L_∞.

Define $\phi(z) = \sqrt{2}\sin(\pi z)$. On the interval $[-v_0, v_0]$, it can be shown that $\phi(z)$ has variation $V_\phi = 2\sqrt{2}\pi v_0$ with resepct to the dictionary of unit step activation functions $\pm\text{step}(z' - t')$, where $\text{step}(z) = \mathbb{I}\{z > 0\}$, or equivalently, variation $\sqrt{2}\pi v_0$ with respect to the dictionary of signum activation functions with shifts $\pm\text{sgn}(z' - t')$, where $\text{sgn}(z) = 2\text{step}(z) - 1$. This can be seen directly from the identity

$$\sin z = \tfrac{v}{2}\int_0^1 \cos(vt)[\text{sgn}(z/v - t) - \text{sgn}(-z/v - t)]dt,$$

for $|z| \leqslant v$. Evaluation of $\int_0^1 |\cos(vt)|dt$ gives the exact value of ϕ with respect to sgn as $\sqrt{2}\pi v_0$ for integer $v = v_0$. Accordingly, $\mathcal{F}_{v_0, v_1, \phi}$ is contained in $\mathcal{F}_{1, \sqrt{2}\pi v_0 v_1, \text{sgn}}$.

Likewise, for the clipped linear function $\text{clip}(z) = \text{sgn}(z)\min\{1, |z|\}$ a similar identity holds:

$$\sin z = z + \tfrac{v^2}{2}\int_0^1 \sin(vt)[\text{clip}(-2z/v - 2t - 1) - \text{clip}(2z/v - 2t - 1)]dt,$$

for $|z| \leqslant v$. The above form arises from integrating

$$\cos w = \cos v - \tfrac{v}{2}\int_0^1 \sin(vt)[\text{sgn}(-w/v - t) + \text{sgn}(w/v - t)]dt,$$

from $w = 0$ to $w = z$. And likewise, evaluation of $\int_0^1 |\sin(vt)|dt$ gives the exact variation of ϕ with respect to the dictionary of clip activation functions $\pm\text{clip}(z' - t')$ as $V_\phi = \sqrt{2}\pi(v_0^2 + 1)$ for integer $v = v_0$. Accordingly, $\mathcal{F}_{v_0, v_1, \phi}$ is contained in $\mathcal{F}_{2, \sqrt{2}\pi(v_0^2+1)v_1, \text{clip}}$ and hence we have the following corollary.

Corollary 1. *Using the same setup and conditions (9) and (10) as in Theorem 1, the minimax risk for the sigmoid classes $\mathcal{F}_{1, \sqrt{2}\pi v_0 v_1, \text{sgn}}$ and $\mathcal{F}_{2, \sqrt{2}\pi(v_0^2+1)v_1, \text{clip}}$ have the same lower bounds (11) and (12) as for $\mathcal{F}_{v_0, v_1, \text{sine}}$.*

IV. IMPLICATIONS FOR POLYNOMIAL NETS

It is also possible to give minimax lower bounds for the function classes $\mathcal{F}_{v_0, v_1, \phi_\ell}$ with activation function ϕ_ℓ equal to the standardized Hermite polynomial $H_\ell/\sqrt{\ell!}$, where $H_\ell(z) = (-1)^\ell e^{\frac{z^2}{2}}\frac{d^\ell}{dz^\ell}e^{-\frac{z^2}{2}}$. As with Theorem 1, this requires a lower bound on d:

$$\frac{d}{v_0^2} > \left(c_8\frac{v_1^2 n}{v_0^2 \log(d/v_0^2)}\right)^{2/v_0^2}. \tag{13}$$

for some constant $c_8 > 0$. Moreover, we also need a growth condition on the order of the polynomial ℓ:

$$\ell > c_9 \log\left(\frac{v_1^2 n}{v_0^2 \log(d/v_0^2)}\right), \tag{14}$$

for some constant $c_9 > 0$. In light of (13), condition (14) is also satisfied if ℓ is at least a constant multiple of $v_0^2 \log(d/v_0^2)$.

Theorem 2. *Consider the model $Y = f(X) + \varepsilon$ for $f \in \mathcal{F}_{v_0, v_1, \phi_\ell}$, where $\varepsilon \sim N(0, 1)$ and $X \sim N(0, I_d)$. If d and ℓ are large enough so that conditions (13) and (14) are satisfied, respectively, then*

$$R_{n,d}(\mathcal{F}_{v_0, v_1, \phi_\ell}) \geqslant c_{10}\left(\frac{v_0^2 v_1^2 \log(d/v_0^2)}{n}\right)^{1/2}, \tag{15}$$

for some universal constant $c_{10} > 0$.

Proof of Theorem 2. By Lemma 1, if $d \geqslant 10$ and $1 \leqslant d' \leqslant d/10$, there exists a subset $\mathcal{C}$ of $\{0, 1\}^d$ with cardinality at least $M := \sqrt{\binom{d}{d'}}$ such that each element has Hamming weight d' and pairs of elements have minumum Hamming distance $d'/5$. Thus, if a and a' belong to this codebook, $|a \cdot a'| \leqslant (9/10)d'$. Choose $d' = v_0^2$ (assuming that v_0^2 is an integer less than d), and form the collection $\mathcal{B} = \{\theta = a/v_0 : a \in \mathcal{C}\}$. Note that each member of $\mathcal{B}$ has unit ℓ_2 norm and ℓ_1 norm v_0. Moreover, the Euclidean inner product between each pair has magnitude bounded by $9/10$. Next, we use the fact that if $X \sim N(0, I_d)$ and θ, θ' have unit ℓ_2 norm, then $\mathbb{E}[\phi_\ell(\theta \cdot X)\phi_\ell(\theta' \cdot X)] = (\theta \cdot \theta')^\ell$. For an enumeration $\theta_1, \ldots, \theta_M$ of $\mathcal{B}$, define a subclass of $\mathcal{F}_{v_0, v_1, H_\ell}$ by

$$\mathcal{F}_0 = \{f_\omega = \tfrac{v_1}{L}\sum_{k=1}^M \omega_k \phi_\ell(\theta_k \cdot x) : \omega \in \mathcal{A}\},$$

where $\mathcal{A}$ is the set from Lemma 1. Moreover, since each θ_k has unit norm, $\|\omega - \omega'\|_1 \geqslant L/5$, and $\|\omega - \omega'\|_1^2 \leqslant 2L\|\omega - \omega'\|_1$,

$$\begin{aligned}
\|f_\omega - f_{\omega'}\|^2 &= \tfrac{v_1^2}{L^2}[\|\omega - \omega'\|_1 + \\
&\quad \sum_{i \neq j}(\omega_i - \omega_i')(\omega_j - \omega_j')(\theta_i \cdot \theta_j)^\ell] \\
&\geqslant \tfrac{v_1^2}{L^2}[\|\omega - \omega'\|_1 - \|\omega - \omega'\|_1^2 (9/10)^\ell] \\
&\geqslant \tfrac{v_1^2}{L^2}\|\omega - \omega'\|_1 (1 - 2L(9/10)^\ell) \\
&\geqslant \tfrac{v_1^2}{L}(1 - 2L(9/10)^\ell) \\
&\geqslant \tfrac{v_1^2}{10L},
\end{aligned}$$

provided $\ell > \frac{\log(4L)}{\log(10/9)}$. A separation of ϵ^2 determines $L = (v_1/(\sqrt{10}\epsilon))^2$. If $L \leqslant \sqrt{M}$, or equivalently, $\epsilon \geqslant v_1 M^{-1/4}$, then $\log(\#\mathcal{F}_0)$ is at least a constant multiple of $\log N_{d>v_0}(\epsilon) = (v_0 v_1/\epsilon)^2 \log(d/v_0^2)$. As before in Theorem 1,

a minimax lower bound ϵ_n^2 on the $L_2(P)$ squared error risk is determined by matching

$$\epsilon_n^2 = \frac{\log N_{d>v_0}(\epsilon_n)}{n},$$

which yields

$$\epsilon_n^2 = \left(\frac{v_0^2 v_1^2 \log(d/v_0^2)}{n}\right)^{1/2}.$$

If conditions (13) and (14) are satisfied, $N_{d>v_0}(\epsilon_n)$ is a valid lower bound on the ϵ_n-packing number of $\mathcal{F}_{v_0,v_1,\phi_\ell}$. $\qquad\square$

V. Discussion

Our risk lower bound of the form $\left(\frac{v_0 v_1^2 \log(1+d/v_0)}{n}\right)^{1/2}$ shows that in the very high-dimensional case, it is the $v_0 v_1^2/n$ to a half-power that controls the rate (to within a logarithmic factor). The v_0 and v_1, as ℓ_1 norms of the inner and outer coefficient vectors, have the interpretations as the effective dimensions of these vectors. Indeed, a vector in $\mathbb{R}^d$ with bounded coefficients that has v_0 non-negligible coordinates has ℓ_1 norm of this order. These rates confirm that it is a power of these effective dimensions over sample size n (instead of the full ambient dimension d) that controls the main behavior of the statistical risk. Our lower bounds on packing numbers complement the upper bound covering numbers in [25] and [13]. Our rates are akin to those obtained in [26] for high-dimensional linear regression. However, there is an important difference. The richness of $\mathcal{F}_{v_0,v_1}$ is largely determined by the sizes of v_0 and v_1 and $\mathcal{F}_{v_0,v_1}$ more flexibly represents a larger class of functions. It would be interesting to see if the gap between the powers $1/2$ and $1/3$ could be closed by improving either the lower bound in (11) or the upper bound in (8).

References

[1] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, no. 1, pp. 115–133, 1994.

[2] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.*, vol. 36, no. 1, pp. 64–94, 2008. [Online]. Available: http://dx.doi.org/10.1214/009053607000000631

[3] C. Huang, G. L. Cheang, and A. R. Barron, "Risk of penalized least squares, greedy selection and ℓ_1 penalization for flexible function libraries," *Yale University, Department of Statistics technical report*, 2008. [Online]. Available: http://www.stat.yale.edu/~arb4/publications_files/RiskGreedySelectionAndL1penalization.pdf

[4] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, no. 3, pp. 301–413, 1999. [Online]. Available: http://dx.doi.org/10.1007/s004400050210

[5] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, no. 4, pp. 1034–1054, 1991. [Online]. Available: http://dx.doi.org/10.1109/18.86996

[6] A. R. Barron, C. Huang, J. Li, and X. Luo, "The MDL principle, penalized likelihoods, and statistical risk," *Workshop on Information Theory Methods in Science and Engineering, Tampere, Finland*, 2008.

[7] ——, "The MDL principle, penalized likelihoods, and statistical risk," *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday, Tampere University Press, Tampere, Finland. Editor Ioan Tabus*, 2008.

[8] A. R. Barron and S. Chatterjee, "Information theoretic validity of penalized likelihood," *Proceedings IEEE International Symposium on Information Theory, Honolulu, HI*, pp. 3027–3031, June 2014.

[9] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, 1992. [Online]. Available: http://dx.doi.org/10.1214/aos/1176348546

[10] A. R. Barron, "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

[11] 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

[12] 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

[13] J. M. Klusowski and A. R. Barron, "Risk bounds for high-dimensional ridge function combinations including neural networks," *arXiv Preprint*, 2017. [Online]. Available: https://arxiv.org/pdf/1607.01434.pdf

[14] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999. [Online]. Available: http://dx.doi.org/10.1214/aos/1017939142

[15] X. Chen and H. White, "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 682–691, 1999. [Online]. Available: http://dx.doi.org/10.1109/18.749011

[16] S. Mendelson, "On the size of convex hulls of small sets," *J. Mach. Learn. Res.*, vol. 2, no. 1, pp. 1–18, 2002. [Online]. Available: http://dx.doi.org/10.1162/153244302760185225

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] I. A. Ibragimov and R. Z. Hasminksii, "Nonparametric regression estimation," *Dokl. Akad. Nauk SSSR*, vol. 252, no. 4, pp. 780–784, 1980.

[19] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," *Z. Wahrsch. Verw. Gebiete*, vol. 65, no. 2, pp. 181–237, 1983. [Online]. Available: http://dx.doi.org/10.1007/BF00532480

[20] ——, "On estimating a density using Hellinger distance and some other strange facts," *Probab. Theory Relat. Fields*, vol. 71, no. 2, pp. 271–291, 1986. [Online]. Available: http://dx.doi.org/10.1007/BF00332312

[21] F. Gao, C.-K. Ing, and Y. Yang, "Metric entropy and sparse linear approximation of ℓ_q-hulls for $0 < q \leqslant 1$," *J. Approx. Theory*, vol. 166, pp. 42–55, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jat.2012.10.002

[22] R. L. Graham and N. J. A. Sloane, "Lower bounds for constant weight codes," *IEEE Trans. Inform. Theory*, vol. 26, no. 1, pp. 37–43, 1980. [Online]. Available: http://dx.doi.org/10.1109/TIT.1980.1056141

[23] A. B. Tsybakov, *Introduction to nonparametric estimation*, ser. Springer Series in Statistics. Springer, New York, 2009, revised and extended from the 2004 French original, Translated by Vladimir Zaiats. [Online]. Available: http://dx.doi.org/10.1007/b13794

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

[25] P. 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*, vol. 44, no. 2, pp. 525–536, 1998. [Online]. Available: http://dx.doi.org/10.1109/18.661502

[26] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6976–6994, 2011. [Online]. Available: http://dx.doi.org/10.1109/TIT.2011.2165799