Generalization Analysis of CNNs for Classification on Spheres

Han Feng, Shuo Huang, and Ding-Xuan Zhou,

***Abstract*—Deep learning based on deep convolutional neural networks (CNNs) is extremely efficient in solving classification problems in speech recognition, computer vision, and many other fields. But there is no enough theoretical understanding about this topic, especially the generalization ability of the induced CNN algorithms. In this paper, we develop some generalization analysis of a deep CNN algorithm for binary classification with data on spheres. An essential property of the classification problem is the lack of continuity or high smoothness of the target function associated with a convex loss function such as the hinge loss. This motivates us to consider the approximation of functions in the L_p space with $1 \le p \le \infty$. We provide rates of L_p-approximation when the approximated function lies in a Sobolev space and then present generalization bounds and learning rates for the excess misclassification error of the deep CNN classification algorithm. Our novel analysis is based on efficient cubature formulae on spheres and other tools from spherical analysis and approximation theory.**

***Index Terms*—deep learning, convolutional neural networks, classification problems, generalization error bounds, spherical analysis.**

I. Introduction

A Binary classification problem with an input (compact metric) space X of instances and output space $Y = \{-1, 1\}$ of two labels aims at learning a (binary) classifier from samples which separates the instances in X into two classes. With a Borel probability measure ρ on $Z := X \times Y$ governing the sampling process, the performance of a classifier $\mathcal{C} : X \to Y$ is assessed by the so-called **misclassification error** defined as the probability of the event $\{(x, y) \in X \times Y : \mathcal{C}(x) \neq y\}$, that is,

$$\mathcal{R}(\mathcal{C}) := \int_{X\times Y} I(-y, \mathcal{C}(x))d\rho = \text{Prob}\{\mathcal{C}(x) \neq y\},$$

where $I(a, b) = 1$ if $a = b$, and 0 otherwise. The best classifier which minimizes the misclassification error is called a **Bayes rule** f_c given by $f_c(x) = 1$ if $\rho(y = 1|x) \ge \rho(y = -1|x)$ and -1 otherwise, where $\rho(\cdot|x)$ denotes the conditional distribution of ρ for $x \in X$. It turns out that for many convex loss functions $\phi : \mathbb{R} \to \mathbb{R}_+$, a Bayes rule can be expressed as $f_c = \text{sgn}(f_\rho^\phi)$ with f_ρ^ϕ being a minimizer of the **generalization error**

$$\mathcal{E}^\phi(f) = \int_Z \phi(yf(x))d\rho$$

over the set of measurable functions $f : X \to \mathbb{R}$. So learning a Bayes rule is reduced to approximating f_ρ^ϕ from hypothesis spaces. A special property of the classification problem is that a Bayes rule f_c taking binary values in $\{1, -1\}$ is often discontinuous, and f_ρ^ϕ may also be discontinuous such as $f_\rho^\phi = f_c$ for the hinge loss $\phi(v) = (1-v)_+ := \max\{1-v, 0\}$. This motivates us to study approximation and learning in L_p spaces with $1 \le p \le \infty$ of functions f_ρ^ϕ in Sobolev spaces W_p^r with small regularity index $r > 0$.

We study binary classification algorithms with hypothesis spaces generated by **deep convolutional neural networks** (CNNs), which is a special kind of deep neural networks produced by convolutions. Classification with CNNs has achieved remarkable successes in many practical applications [32], [19], [10]. Theoretical verifications, however, still lack much. In [37], the uniform approximation by CNNs of functions with Hölder regularity index $r > \frac{d}{2} + 2$ on domains in $\mathbb{R}^d$ was considered by the last author. In [13], we analyzed the uniform approximation of nonsmooth functions with small Hölder regularity index $r > 0$ on the unit sphere $X = \mathbb{S}^{d-1}$ of $\mathbb{R}^d$ which is the input space throughout this paper. See comparisons in Table I. Spherical data arise naturally in many fields such as cosmic microwave background analysis [12], global ionospheric prediction to geomagnetic storms [20], climate change modelling, environmental governance, meteorology, remote sensing and other spherical signals. In this paper, we carry out not only analysis for the L_p approximation by deep CNNs with $p < \infty$ but also improve our previous results in the case of uniform approximation with $p = \infty$. This is achieved by our novel idea of using methods from spherical analysis and approximation theory. Our quantitative estimates for the L_p approximation lead to generalization bounds for the excess misclassification error.

The convolution of a sequence w on $\mathbb{Z}$ supported in $\{0, \cdots, S\}$ for some filter length $S \in \mathbb{N}$ with another sequence supported in $\{1, \cdots, D\}$, regarded as a vector $v = (v_1, \ldots, v_D)$, is defined as a sequence $w * v$ given by

$$(w * v)_i = \sum_{k\in\mathbb{Z}} w_{i-k}v_k = \sum_{k=1}^{D} w_{i-k}v_k, \qquad i \in \mathbb{Z},$$

which is supported in $\{1, \cdots, D + S\}$. For an input $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, a deep CNN with J hidden layers of neurons $\{h^{(j)} : \mathbb{R}^d \to \mathbb{R}^{d_j}\}$ and widths $\{d_j := d + js\}$ is defined in [37], [39] by $h^{(0)}(x) = x$ and iteratively

$$h^{(j)}(x) = \sigma\left(T^{(j)}\ h^{(j-1)}(x) - b^{(j)}\right), \quad j = 1, \ldots, J, \quad (1)$$

where $T^{(j)} := T^{w^{(j)}}$ is a Toeplitz type **convolutional matrix**

Han Feng is with the Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (e-mail: hanfeng@cityu.edu.hk).

Shuo Huang is with the Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (e-mail: shuang56-c@my.cityu.edu.hk).

Ding-Xuan Zhou is with School of Data Science, Department of Mathematics, and Liu Bie Ju Centre for Mathematical Sciences, City University of Hong Kong, Kowloon, Hong Kong (e-mail: mazhou@cityu.edu.hk).

given with a filter sequence $w = w^{(j)}$ and $D = d_{j-1}$, $D+S = d_j$ by

$$T^w = \begin{bmatrix} w_0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ w_1 & w_0 & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ w_S & w_{S-1} & \cdots & w_0 & 0 & \cdots & 0 \\ 0 & w_S & \cdots & w_1 & w_0 & 0\cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w_S & \cdots & w_1 & w_0 \\ 0 & \cdots & 0 & 0 & w_S & \cdots & w_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & w_S \end{bmatrix} \tag{2}$$

with dimension $\mathbb{R}^{(D+S)\times D}$. Throughout the paper we take the rectified linear unit (ReLU) activation function

$$\sigma(u) = \max\{u, 0\}, \qquad u \in \mathbb{R}.$$

After the last CNN layer, we apply a **downsampling** operator $\mathfrak{D}_d$ introduced in [38] for $v = (v_i)_{i=1}^D \in \mathbb{R}^D$ defined by $\mathfrak{D}_d(v) = (v_{id})_{i=1}^{\lfloor D/d \rfloor}$, where $\lfloor u \rfloor$ denotes the integer part of $u > 0$. Then we add two fully connected layers $h^{(J+1)}, h^{(J+2)}$ with widths $\mathcal{D}_1, \mathcal{D}_2 > 0$, respectively, connection matrices $F^{(J+1)}, F^{(J+2)}$ and bias vectors $b^{(J+1)}, b^{(J+2)}$, to be determined. Precisely,

$$h^{(J+1)}(x) = \sigma\left(F^{(J+1)}\mathfrak{D}_d\left(h^{(J)}(x)\right) - b^{(J+1)}\right) \tag{3}$$

and

$$h^{(J+2)}(x) = \sigma\left(F^{(J+2)}h^{(J+1)}(x) - b^{(J+2)}\right). \tag{4}$$

Such a network with many convolutional layers followed by downsampling operations and very few fully connected layers is quite common in practical applications [19], [14]. An output function of our network takes the form $c^{(J+2)} \cdot h^{(J+2)}(x) - A : X \to \mathbb{R}$ with a coefficient vector $c^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}$ and a bias $A \in \mathbb{R}$.

II. Main Results

In this paper we conduct generalization analysis of deep CNNs used for binary classification with data on spheres. When the hinge loss is used, the target function $f_\rho^\phi = f_c$ is a binary function which is not suitable for uniform approximation.

A. Approximation by deep CNNs in L_p spaces

Our first main result is the following estimate for the approximation ability of our CNN network with respect to the p-norm $\|\cdot\|_p$ in the L_p space to approximate functions from the Sobolev space $W_p^r(\mathbb{S}^{d-1})$ with $1 \le p \le \infty, r > 0$, to be defined in Section III.

Theorem 1: Let $d \ge 3$, $2 \le S \le d$, $r > 0$ and $1 \le p \le \infty$. Then there exists a constant $\hat{c}_d \ge d$ depending only on d such that for any $J \ge \frac{\hat{c}_d}{S-1}$ and $f \in W_p^r(\mathbb{S}^{d-1})$, a deep neural network consisting of J layers of CNNs and two fully connected layers of widths $\mathcal{D}_1 = (2N+3)\lfloor (d+JS)/d \rfloor, \mathcal{D}_2 = \lfloor (d+JS)/d \rfloor$ respectively and $N = \left\lceil \left(\frac{(S-1)J}{\hat{c}_d}\right)^{\frac{d+3+r}{2(d-1)}} \right\rceil$ produces an output function $c^{(J+2)} \cdot h^{(J+2)}(x) - A : X \to \mathbb{R}$ with $c^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}$ and $A \in \mathbb{R}$ satisfying

$$\left\| f - c^{(J+2)} \cdot h^{(J+2)}(x) + A \right\|_p \le C\|f\|_{W_p^r(\mathbb{S}^{d-1})} J^{-\frac{r}{d-1}}, \tag{5}$$

where C is a constant depending only on d, r, S and p. The total number of free parameters $\mathcal{N}$ in the network can be bounded as

$$\mathcal{N} \le (3S+5)J^{\max\left\{1, \frac{d+3+r}{2(d-1)}\right\}} + 5.$$

The proof of Theorem 1 will be given in Section III. More details on the explicit structure of our neural network can be found in Appendix A.

The approximation order in Theorem 1 might be improved when features of the approximated functions other than the regularity are used. Such features may be found for functions on spheres due to the spherical properties. For example, a special class of functions on spheres consists of spherical polynomials to be defined later. We observe that for these functions, better approximation can be achieved.

Corollary 1: Let $d \ge 3$, $2 \le S \le d$, $r > 0$, $1 \le p \le \infty$ and $n, N \in \mathbb{N}$. Then there exists a constant $\hat{c}_d \ge d$ depending only on d such that for any spherical polynomial f of degree n, $l \in \mathbb{N}$ and $J \ge \frac{\hat{c}_d n^{d-1}}{S-1}$, a deep neural network consisting of J layers of CNNs and two fully connected layers of widths $\mathcal{D}_1 = (2N+3)\lfloor (d+JS)/d \rfloor, \mathcal{D}_2 = \lfloor (d+JS)/d \rfloor$ respectively produces an output function $c^{(J+2)} \cdot h^{(J+2)} - A$ with $c^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}$ and $A \in \mathbb{R}$ satisfying

$$\left\| f - c^{(J+2)} \cdot h^{(J+2)} + A \right\|_p \le Cn^{d+3}\|f\|_p N^{-2}, \tag{6}$$

where C is a constant depending only on d, r, S and p, and

$$\mathcal{N} \le J(3S+2) + m + 2N + 4$$

is the total number of free parameters in the network.

B. Generalization error bounds for classification

Once we understand the approximation ability of the deep network, we can carry out generalization analysis of empirical risk minimization (ERM) algorithms implemented over hypothesis spaces induced by the network. Here we restrict the parameters to have a uniform bound $R > 0$ and take the hypothesis space of functions on $\mathbb{S}^{d-1}$ induced by our network stated in Theorem 1 as

$$\begin{aligned} \mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S,R} = \Big\{ & c^{(J+2)} \cdot h^{(J+2)}(x) - A : \ c^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}, \\ & A \in \mathbb{R}, \|w^{(j)}\|_\infty, \|b^{(j)}\|_\infty, \|F^{(J+1)}\|_\infty, \|F^{(J+2)}\|_\infty, \\ & \|c^{(J+2)}\|_\infty \le R \Big\}. \end{aligned} \tag{7}$$

The classification algorithm we study produces a classifier $\operatorname{sgn}(\widehat{f}_{\mathbf{z}})$ as the sign of $\widehat{f}_{\mathbf{z}}$, a minimizer over the hypothesis space $\mathcal{H} = \mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S,R}$ of the empirical error $\mathcal{E}_{\mathbf{z}}(f)$ associated with a convex loss function ϕ and a random sample $\mathbf{z} := \{(x_i, y_i)\}_{i=1}^{\mathcal{M}}$ drawn according to ρ

$$\widehat{f}_{\mathbf{z}} := \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f) = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \phi(y_i f(x_i)) \right\}. \tag{8}$$

Our target is to estimate the excess misclassification error

$$\mathcal{R}(\mathrm{sgn}(\widehat{f_{\mathbf{z}}})) - \mathcal{R}(f_c). \tag{9}$$

Its convergence rates depend on the convexity of the loss ϕ and noise level of the underlying distribution ρ, which can be measured simultaneously by the **variancing power** [33] of the pair (ϕ, ρ) defined as the maximum $\tau \in [0,1]$ such that for some $C_1 > 0$,

$$\mathbb{E}\left\{\left(\phi(yf(x)) - \phi\left(yf_\rho^\phi(x)\right)\right)^2\right\} \le C_1\left(\mathcal{E}(f) - \mathcal{E}\left(f_\rho^\phi\right)\right)^\tau \tag{10}$$

holds for any measurable function $f : X \to \mathbb{R}$. A generalization error bound with a general loss function ϕ will be given in Theorem 5 of Section IV. As an illustration, we state learning rates for the p-norm loss $\phi(v) = (1-v)_+^p$ with $p > 1$ and the hinge loss with $p = 1$. This is our second main result, to be proved in Section IV, with learning rates of the CNN classification algorithm given in terms of the approximation error

$$\mathcal{D}(\mathcal{H}) = \inf_{f \in \mathcal{H}} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi)\right\}.$$

Theorem 2: Let $d \ge 3$, $2 \le S \le d$, $r > 0$, $1 \le p < \infty$ and $\phi(v) = (1-v)_+^p$. If the pair (ϕ, ρ) has a variancing power $\tau \in [0,1]$ with (10) valid and the approximation error of the CNN hypothesis space $\mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S,R}$ satisfies

$$\mathcal{D}(\mathcal{H}) = \inf_{f \in \mathcal{H}} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi)\right\} \le C_0 J^{-\frac{rp}{d-1}} \tag{11}$$

with a constant C_0 independent of the depth J, then by choosing

$$J = \begin{cases} \left\lceil (\mathcal{M}/\log\mathcal{M})^{\frac{d-1}{(\beta+1)(d-1)+pr(2-\tau)}} \right\rceil, & \text{if } p > 1, \\ \left\lceil (\mathcal{M}/\log\mathcal{M})^{\frac{d-1}{2\beta(d-1)+r(2-\tau)}} \right\rceil, & \text{if } p = 1, \end{cases} \tag{12}$$

for any $\delta > 0$, with confidence $1-\delta$, the excess misclassification error $\mathcal{R}(\mathrm{sgn}(\widehat{f_{\mathbf{z}}})) - \mathcal{R}(f_c)$ of the induced classifier $\mathrm{sgn}(\widehat{f_{\mathbf{z}}})$ can be bounded as

$$\begin{cases} \widetilde{C}\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}} \log\frac{2}{\delta}, & \text{if } p > 1, \\ \widetilde{C}\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{r}{2\beta(d-1)+r(2-\tau)}} \log\frac{2}{\delta}, & \text{if } p = 1, \end{cases} \tag{13}$$

where $\beta = \max\left\{1, \frac{d+3+r}{2(d-1)}\right\}$ and $\widetilde{C}$ is a constant independent of $\mathcal{M}$ or δ.

Based on Theorem 1 and bounds in [7] for $\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi)$ in terms of $\|f - f_\rho^\phi\|_{L^p_{\rho_X}}^p$ with respect to the marginal distribution ρ_X on X, we know that the decay (11) of the approximation error is a reasonable assumption.

The power indices in (13) and the learning rates can be better demonstrated when r is large enough. In this case, the learning rates can be of order $O\left((\log\mathcal{M}/\mathcal{M})^{\frac{1}{2-\tau}-\epsilon}\right)$ for $p = 1$, where $\epsilon > 0$ can be arbitrarily small.

C. Improved learning rates under noise conditions

The variancing power τ and hence learning rates can be explicitly found when some noise conditions are imposed. Here we consider the **Tsybakov noise condition** [31] with an exponent $\theta > 0$ which is defined for a Borel probability measure ρ to satisfy

$$\rho_X\left(\{x \in X : 0 < |f_\rho(x)| \le c_\theta t\}\right) \le t^\theta, \quad \forall t > 0 \tag{14}$$

with a positive constant c_θ where $f_\rho : X \to \mathbb{R}$ is the regression function given by $f_\rho(x) = \int_Y y d\rho(y|x) = \rho(y=1|x) - \rho(y=-1|x)$ for $x \in X$ or equivalently $f_\rho(x) := \eta(x) - (1-\eta(x)) = 2\eta(x) - 1$ with $\eta(x) := \rho(y=1|x)$ called the **conditional class probability**. Our last main result, to be proved in Section V, present improved learning rates when such a noise condition is imposed and the 2-norm loss is used for classification.

Theorem 3: Let $d \ge 3$, $2 \le S \le d$, and $\phi(v) = (1-v)_+^2$. If condition (11) for the decay of the approximation error is valid with $p = 2$ and some $r > 0$, and the Tsybakov noise condition (14) is satisfied for some $\theta > 0$, then by taking

$$J = \left\lceil (\mathcal{M}/\log\mathcal{M})^{\frac{d-1}{(\beta+1)(d-1)+2r}} \right\rceil,$$

for any $\delta > 0$, with confidence $1-\delta$, there holds

$$\mathcal{R}(\mathrm{sgn}(\widehat{f_{\mathbf{z}}})) - \mathcal{R}(f_c) \le \widetilde{C}\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{2r\theta}{(2+\theta)((\beta+1)(d-1)+2r)}} \log\frac{2}{\delta},$$

where $\widetilde{C}$ is a constant independent of $\mathcal{M}$ or δ.

When the indices r for the approximation error and θ for the noise are large enough, the above learning rate can be of order $O\left((\log\mathcal{M}/\mathcal{M})^{4/5-\epsilon}\right)$, where $\epsilon > 0$ can be arbitrarily small. This rate verifies the efficiency of CNN algorithms in solving classification tasks. In the special case when the target function is a spherical polynomial, the learning rate can be further improved.

Theorem 4: Let $d \ge 3$, $2 \le S \le d$, and $\phi(v) = (1-v)_+^2$. Let f_ρ be a spherical polynomial. Under the Tsybakov noise condition (14) for some $\theta > 0$, we have with confidence $1-\delta$,

$$\mathcal{R}(\mathrm{sgn}(\hat{f}_{\mathbf{z}})) - \mathcal{R}(f_c) \le C\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{4\theta/5(2+\theta)} \log\frac{2}{\delta}.$$

D. Discussion

The classical topic of approximating functions by shallow or multi-layer neural networks was well developed 30 years ago when the networks are **fully connected** meaning that the connections in (1) are full matrices instead of sparse convolutional ones expressed by (2) in this paper. The fully connectedness yields nice approximation rates [2], [25] when the activation function is a C^∞ sigmoid type function. For example, with a localized Taylor expansion approach, for approximating $f \in W_\infty^r([-1,1]^d)$, convergence rates of order $O(N^{-r/d})$ with a shallow network of N hidden neurons were obtained in [25] if for some $b \in \mathbb{R}$ and some integer $\ell \in \mathbb{N} \setminus \{1\}$, the C^∞ activation function σ satisfies $\sigma^{(k)}(b) \ne 0$ for all $k \in \mathbb{Z}_+$ and $\lim_{u\to-\infty} \sigma(u)/|u|^\ell = 0$ and $\lim_{u\to\infty} \sigma(u)/u^\ell = 1$. The problem for ReLU which does not satisfy these extra conditions was solved recently in [18], [34], [4], [27], [28], [30] for fully connected networks.

Deep CNNs have different structures induced by convolutions. Their approximation theory was recently developed in [37] for universality, in [38] for comparisons with fully

connected networks, in [13] for analysis with spherical data, in [24] for demonstrating superiority in approximating radial functions, and in [40] for CNNs induced by 2-D convolutions. ResNet-type CNNs were studied in [26].

TABLE I
APPROXIMATION ERROR USING DEEP CNNS

Regularity	Range	Error rate
$f \in W_2^r(\mathbb{R}^d)$	$r > 2 + \frac{d}{2}$	$O(J^{-(\frac{1}{2}+\frac{1}{d})})$, [37, Theorem 2]
$f \in W_\infty^r(\mathbb{S}^{d-1})$	$r > 0$	$O(J^{-\min\{\frac{r}{2(d-1+\tau)}, \frac{1}{2}\}})$, [13, Theorem 1]
$f \in W_p^r(\mathbb{S}^{d-1})$	$r > 0$, $1 \le p \le \infty$	$O(J^{-\frac{r}{d-1}})$, [Theorem 1 here]

The problem of L_p approximation with $p < \infty$ appears naturally in the setting of binary classification. The approximated function here is often discontinuous and may be assumed to be in a Sobolev space $W_p^r(X)$ with small $r > 0$, which involves the topic of approximating non-smooth functions studied in [16].

Convergence rates of the excess misclassification error with the $0-1$ loss were well studied two decades ago. Convergence rates of order $O(\mathcal{M}^{-1/2})$ can be attained using oracle inequalities [23], [35]. It was shown in [31] that the minimax lower bound is $O(\mathcal{M}^{-\alpha(\theta+1)/\{\alpha(\theta+2)+(d-1)\theta\}})$ when the empirical risk minimizer is taken over all measurable classifiers and the decision boundaries are generated by α-Hölder smooth functions with Tsybakov noise condition of exponent θ. Furthermore, fast rates of order $O(\mathcal{M}^{\epsilon-1})$ with an arbitrarily small $\epsilon > 0$ can be attained when α and θ are large enough. As for using neural networks, minimax optimal rates $O(\mathcal{M}^{-\alpha(\theta+1)/\{\alpha(\theta+2)+(d-1)\theta\}})$ and $O(\mathcal{M}^{\alpha(\theta+1)/\{\alpha(\theta+2)+d\}})$ under both Tsybakov noise condition and an additional condition for α-Hölder smooth decision boundary or α-Hölder smooth conditional class probability $\eta(x)$ respectively, are shown in [17] when the empirical risk minimizer $\widehat{f}_{\mathbf{z}}$ is generated by deep neural networks (DNN) with specified structures. To the best of our knowledge, no misclassification error rate has been established for deep CNN classifiers. The related results can be seen in Table II.

Choosing the depth J is a crucial topic in deep learning. While increasing the depth can reduce the approximation error, it makes the sample error larger. Our choice (12) of the depth J makes a trade-off between approximation error and sample error. This phenomenon of balancing the approximation ability and capacity has been practically observed in a large literature and theoretically verified for ReLU fully connected networks in [8], [9], [15].

III. ESTIMATES FOR APPROXIMATION BY DEEP CNNS

In this section, we establish error estimates for the approximation of $f_\rho^\phi \in W_p^r(\mathbb{S}^{d-1})$ by our deep CNN network. Such a result was obtained for the case $p = \infty$ in our previous work [13]. Here we not only extend the estimate to the case $1 \le p < \infty$ but also improve the convergent rates. The key novelty of our significantly improved analysis is to apply an efficient cubature formula on spheres and a tighter spline interpolation. So ideas and methods from spherical analysis and approximation theory play essential roles in our estimates. In particular, the following preliminaries are needed.

TABLE II
EXCESS MISCLASSIFICATION ERROR

Hypothesis space	Loss	Condition	Rate
Measurable functions	$0-1$ loss	θ-noise condition; α-Hölder decision boundary	$O(\mathcal{M}^{-\frac{\alpha(\theta+1)}{\alpha(\theta+2)+(d-1)\theta}})$ [31, Theorem 1]
DNN	hinge		$O(\mathcal{M}^{-\frac{\alpha(\theta+1)}{\alpha(\theta+2)+(d-1)(\theta+1)}})$ [17, Theorem 1]
Deep CNNs	$1-$norm	$f_\rho^\phi \in W_p^r(\mathbb{S}^{d-1})$	$O(\mathcal{M}^{-\frac{r}{2\beta(d-1)+r(2-\tau)}})$ [Theorem 2 here]
	$p-$norm		$O(\mathcal{M}^{-\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}})$ [Theorem 2 here]
	$2-$norm	$f_\rho^\phi \in W_p^r(\mathbb{S}^{d-1})$; θ-noise condition	$O(\mathcal{M}^{-\frac{2r\theta}{(2+\theta)((\beta+1)(d-1)+2r)}})$ [Theorem 3 here]

A. Spherical harmonics and cubature formulae

A spherical harmonic of degree n on the sphere $\mathbb{S}^{d-1}$ is a homogeneous polynomial P_n^d of degree n defined on $\mathbb{R}^d$ satisfying $\Delta P_n^d = 0$, where Δ is the Laplace operator on $\mathbb{R}^d$. Denote $\mathcal{H}_n^d$ as the set of all spherical harmonics of degree n on $\mathbb{S}^{d-1}$. Its dimension is

$$N(n,d) = \binom{n+d-1}{n} - \binom{n+d-3}{n-2} \le C_d\, n^{d-2}, \quad (15)$$

where $C_d > 0$ is a constant depending only on d.

The spaces $\mathcal{H}_n^d$ of spherical harmonics can also be characterized as eigenfunction spaces of the Laplace-Beltrami operator Δ_0 on $\mathbb{S}^{d-1}$, that is,

$$\mathcal{H}_n^d = \{f \in C^2(\mathbb{S}^{d-1}) : \Delta_0 f = -\lambda_n f\},$$

where $\lambda_n = n(n+d-2)$ and $C^2(\mathbb{S}^{d-1})$ denotes the space of all twice continuously differentiable functions on $\mathbb{S}^{d-1}$. We define the Sobolev space $W_p^r(\mathbb{S}^{d-1})$ to be a subspace of $L_p(\mathbb{S}^{d-1})$, $1 \le p \le \infty$, $r > 0$, with the finite norm

$$\|f\|_{W_p^r(\mathbb{S}^{d-1})} := \left\|(-\Delta_0 + I)^{r/2} f\right\|_{L_p(\mathbb{S}^{d-1})}, \quad (16)$$

where $\|g\|_{L_p(\mathbb{S}^{d-1})} = \left(\int_{\mathbb{S}^{d-1}} |g(x)|^p d\mu\right)^{1/p}$ denotes the L_p norm with respect to the normalized spherical measure μ on $\mathbb{S}^{d-1}$. For $d \ge 3$, let $C_n^\lambda(t)$ be the Gegenbauer polynomial of degree n with parameter $\lambda := \frac{d-2}{2}$. It is well known that $L_2(\mathbb{S}^{d-1})$ can be orthogonally decomposed as

$$L_2(\mathbb{S}^{d-1}) = \bigoplus_{n=0}^{\infty} \mathcal{H}_n^d$$

and for any $x, y \in \mathbb{S}^{d-1}$, $\frac{n+\lambda}{\lambda} C_n^\lambda(\langle x, y\rangle)$ is a reproducing kernel of $\mathcal{H}_n^d$ in the sense that

$$\int_{\mathbb{S}^{d-1}} p(y) \frac{n+\lambda}{\lambda} C_n^\lambda(\langle x, y\rangle) d\mu(y) = p(x), \quad \forall p \in \mathcal{H}_n^d, \quad (17)$$

where $\langle x, y\rangle$ is the inner product on $\mathbb{R}^d$.

Given a smooth function $\eta \in C^\infty([0,\infty))$ with $\eta(t) = 1$ for $t \in [0,1]$, $0 \le \eta(t) \le 1$ for $t \in [1,2]$ and $\eta(t) = 0$ for $t \ge 2$, we set

$$K_n(t) = \sum_{\ell=0}^{2n} \left[\eta(\frac{\ell}{n})\right]^2 \frac{\lambda+\ell}{\lambda} C_\ell^\lambda(t)$$

and define a linear operator $L_n : L_p(\mathbb{S}^{d-1}) \to L_p(\mathbb{S}^{d-1})$ for $f \in L_p(\mathbb{S}^{d-1})$ by

$$L_n(f)(x) = \int_{\mathbb{S}^{d-1}} f(y) K_n(\langle x, y\rangle) d\mu(y), \qquad x \in \mathbb{S}^{d-1}.$$

This integral linear operator is bounded and provides good approximations of Sobolev functions.

Lemma 1: For $n \in \mathbb{N}$, $r > 0$, $1 \le p \le \infty$ and $f \in W_p^r\left(\mathbb{S}^{d-1}\right)$, there holds

$$\|f - L_n(f)\|_{L_p(\mathbb{S}^{d-1})} \le c_1 2^{d-1} n^{-r} \|f\|_{W_p^r(\mathbb{S}^{d-1})}, \tag{18}$$

where c_1 is a constant depending only on the function η. Furthermore, with a constant $C_2 > 0$ depending on d and η,

$$\|L_n(f)\|_{L_p(\mathbb{S}^{d-1})} \le C_2 \|f\|_{L_p(\mathbb{S}^{d-1})}. \tag{19}$$

To get a discrete representation of $L_n(f)$, we shall use a cubature formula for integration of polynomials of degree $4n$ on $\mathbb{S}^{d-1}$, $d \ge 3$, see [6, Theorem 3.1].

Lemma 2: There exists a constant $c_2 > 0$ depending only on d such that for any $m \ge c_2 n^{d-1}$, there exist positive numbers λ_j and points $z_j \in \mathbb{S}^{d-1}$, $j = 1, \dots, m$, satisfying

$$\int_{\mathbb{S}^{d-1}} f(x) \mathrm{d}\mu(x) = \sum_{j=1}^m \lambda_j f(z_j), \quad \forall f \in \Pi_{4n}\left(\mathbb{S}^{d-1}\right),$$

where $\Pi_{4n}\left(\mathbb{S}^{d-1}\right)$ denotes the space of polynomials of degree up to $4n$ on $\mathbb{S}^{d-1}$. Furthermore, for $f \in \Pi_{4n}(\mathbb{S}^{d-1})$,

$$\|f\|_p \asymp \begin{cases} \left(\sum_{j=1}^m \lambda_j |f(z_j)|^p\right)^{\frac{1}{p}}, & \text{if } 1 \le p < \infty, \\ \max_{j=1,\dots,m} n^{d-1} \lambda_j |f(z_j)|, & \text{if } p = \infty, \end{cases} \tag{20}$$

where $A \asymp B$ means there are $c_3, c_4 > 0$ independent of n or m such that $c_3 A \le B \le c_4 A$. Particularly, we say such a family $\{(\lambda_j, z_j)\}_{j=1}^m$ follows a cubature rule of degree $4n$.

B. Proof of Theorem 1

Now we are in a position to prove Theorem 1 based on the following lemmas. Define

$$\tilde{K}_n(t) = \sum_{\ell=0}^{2n} \eta(\frac{\ell}{n}) \frac{\lambda+\ell}{\lambda} C_\ell^\lambda(t),$$

which is a polynomial of degree $2n$.

Lemma 3: Let $d \ge 3$, $r > 0$, $1 \le p \le \infty$. There exist $\lambda_j \in \mathbb{R}$ and $z_j \in \mathbb{S}^{d-1}$, $j = 1, 2, \dots, m$, with $m = \lfloor c_2 + 1 \rfloor n^{d-1}$, such that for any $f \in W_p^r(\mathbb{S}^{d-1})$,

$$\begin{aligned} &\left\| f - \sum_{j=1}^m \lambda_j \alpha_n(f)(z_j) \tilde{K}_n(\langle z_j, \cdot\rangle) \right\|_p \\ &\le c_1 2^{d-1} n^{-r} \|f\|_{W_p^r(\mathbb{S}^{d-1})}, \end{aligned} \tag{21}$$

where

$$\alpha_n(f)(z_j) = \int_{\mathbb{S}^{d-1}} f(y) \tilde{K}_n(\langle z_j, y\rangle) d\mu(y)$$

and $\{(\lambda_j, z_j)\}_{j=1}^m$ follows a cubature rule of degree $4n$.

Proof: By the reproducing property (17) and orthogonal property of $\mathcal{H}_n^d$, we have that for any $x, y \in \mathbb{S}^{d-1}$,

$$\begin{aligned} K_n(\langle x, y\rangle) &= \sum_{\ell=0}^{2n} \eta(\frac{\ell}{n}) \frac{\lambda+\ell}{\lambda} \int_{\mathbb{S}^{d-1}} C_\ell^\lambda(\langle x, z\rangle) \\ &\quad \cdot \sum_{k=0}^{2n} \eta(\frac{k}{n}) \frac{\lambda+k}{\lambda} C_k^\lambda(\langle z, y\rangle) d\mu(z) \\ &= \int_{\mathbb{S}^{d-1}} \tilde{K}_n(\langle x, z\rangle) \tilde{K}_n(\langle z, y\rangle) d\mu(z). \end{aligned} \tag{22}$$

According to Lemma 2, for $c_2 n^{d-1} < m = \lfloor c_2 + 1 \rfloor n^{d-1} \le (c_2+1) n^{d-1}$ there exists a cubature rule $\{(\lambda_j, z_j)\}_{j=1}^m$ of degree $4n$ satisfying

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \tilde{K}_n(\langle x, z\rangle) \tilde{K}_n(\langle z, y\rangle) d\mu(z) \\ &= \sum_{j=1}^m \lambda_j \tilde{K}_n(\langle x, z_j\rangle) \tilde{K}_n(\langle z_j, y\rangle), \end{aligned}$$

which, combining (22), yields that

$$\begin{aligned} L_n(f)(x) &= \int_{\mathbb{S}^{d-1}} f(y) \sum_{j=1}^m \lambda_j \tilde{K}_n(\langle x, z_j\rangle) \tilde{K}_n(\langle z_j, y\rangle) d\mu(y) \\ &= \sum_{j=1}^m \lambda_j \int_{\mathbb{S}^{d-1}} f(y) \tilde{K}_n(\langle z_j, y\rangle) d\mu(y) \tilde{K}_n(\langle x, z_j\rangle). \end{aligned}$$

This completes the proof by using (18). ∎

Proof of Theorem 1: Note that $\tilde{K}_n$ is a polynomial of degree at most $2n$ on $[-1,1]$, by Markov's inequality,

$$\|\tilde{K}_n''\|_\infty \le (2n)^2 \|\tilde{K}_n'\|_\infty \le (2n)^4 \|\tilde{K}_n\|_\infty \le 3^{d+2} n^{d+3}.$$

Here we have used the fact that for $\lambda = \frac{d-2}{2} > 0$,

$$\|C_\ell^\lambda\|_\infty = C_\ell^\lambda(1) = \binom{\ell+d-3}{\ell} \le (\ell+1)^{d-3}$$

holds for all $\ell \in \mathbb{N}$, which implies that $\|\tilde{K}_n\|_\infty \le 5 \cdot 3^{d-2} n^{d-1}$.

By [5, Theorem 2.1], for $N \in \mathbb{N}$, the quasi-interpolant Q_N induced by the hat function $\psi(t) = N(\sigma(t - \frac{1}{N}) - 2\sigma(t) + \sigma(t + \frac{1}{N}))$ defined for continuous functions $g \in C[-1,1]$ by

$$Q_N(g)(t) = \sum_{\ell=-N}^N g(\frac{\ell}{N}) \psi(t - \frac{\ell}{N}), \qquad t \in [-1,1]$$

satisfies

$$\|g - Q_N(g)\|_\infty \le c_5 \frac{1}{N} \omega\left(g', 1/N\right), \qquad g \in C^1[-1,1],$$

where c_5 is an absolute constant and $\omega\left(g', 1/N\right)$ is the modulus of continuity of g' given by

$$\omega(g', 1/N) = \sup_{\substack{u \in [-1, 1-t], \\ t \in [0, 1/N]}} \left\{ |g'(u) - g'(u+t)| \right\}.$$

In particular, for the function $\tilde{K}_n \in C^2[-1,1]$, we have

$$\left\|\tilde{K}_n - Q_N(\tilde{K}_n)\right\|_\infty \le \frac{c_5\|\tilde{K}_n''\|_\infty}{N^2}.$$

Set

$$\mathcal{A}_\ell(\tilde{K}_n) = N\Big(\tilde{K}_n(\frac{\ell-1}{N}) - 2\tilde{K}_n(\frac{\ell}{N}) + \tilde{K}_n(\frac{\ell+1}{N})\Big)$$

for $\ell = -N-1,\dots,N+1$, where we denote $\tilde{K}_n(\frac{j}{N}) = 0$ for $j \notin [-N,\dots,N]$. If we denote

$$\mathcal{B}(t) := \tilde{K}_n(t) - \sum_{\ell=-N-1}^{N+1} \mathcal{A}_\ell(\tilde{K}_n)\sigma(t-\frac{\ell}{N}), \qquad t\in[-1,1],$$

then the above analysis yields

$$\|\mathcal{B}\|_\infty \le \frac{c_5 3^{d+2} n^{d+3}}{N^2}. \tag{23}$$

As shown in the appendix, we can construct our CNN network with parameters given explicitly and the final output given by (45) as

$$\begin{aligned} &c^{(J+2)}\cdot h^{(J+2)}(x) - A \\ &= \sum_{j=1}^m \lambda_j \alpha_n(f)(z_j)\left[\sum_{\ell=-N-1}^{N+1} \mathcal{A}_\ell(\tilde{K}_n)\sigma\left(\langle z_j,x\rangle - \frac{\ell}{N}\right)\right]. \end{aligned}$$

Then from the definition of the function $\mathcal{B}$ and its bound (23), we see by Hölder's inequality that for $1\le p<\infty$ with $p' = \frac{p}{p-1}$,

$$\begin{aligned} &\left\|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\tilde{K}_n(\langle z_j,\cdot\rangle) - c^{(J+2)}\cdot h^{(J+2)}(x) + A\right\|_p^p \\ &= \int_{\mathbb{S}^{d-1}} \left|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\mathcal{B}(\langle z_j,x\rangle)\right|^p d\mu(x) \\ &\le \sum_{j=1}^m \lambda_j \left|\alpha_n(f)(z_j)\right|^p \int_{\mathbb{S}^{d-1}} \left(\sum_{j=1}^N \lambda_j \left|\mathcal{B}(\langle z_j,x\rangle)\right|^{p'}\right)^{\frac{p}{p'}} d\mu(x) \\ &\le \sum_{j=1}^m \lambda_j \left|\alpha_n(f)(z_j)\right|^p \left(\sum_{j=1}^m \lambda_j\right)^{\frac{p}{p'}} \|\mathcal{B}\|_\infty^p. \end{aligned}$$

Note that the reproducing kernel $\frac{k+\lambda}{\lambda}C_k^\lambda(\langle x,y\rangle)$ of $\mathcal{H}_k^d$ can be expressed in terms of an orthonormal basis $\{Y_{\ell,k}\}_{\ell=1}^{N(k,d)}$ of $\mathcal{H}_k^d$ as $\sum_{\ell=1}^{N(k,d)} Y_{\ell,k}(x)Y_{\ell,k}(y)$. So for $z\in\mathbb{S}^{d-1}$,

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} f(y)\tilde{K}_n(\langle y,z\rangle)d\mu(y) \\ &= \sum_{k=0}^{2n}\eta(\frac{k}{n}) \sum_{\ell=1}^{N(k,d)} Y_{\ell,k}(z)\int_{\mathbb{S}^{d-1}} f(y)Y_{\ell,k}(y)d\mu(y), \end{aligned}$$

which is a polynomial of degree $2n$ in z. By (20) in Lemma 2 and a similar bound as (19) in Lemma 1,

$$\begin{aligned} &\sum_{j=1}^m \lambda_j|\alpha_n(f)(z_j)|^p \\ &\le c_3^p \int_{\mathbb{S}^{d-1}} \left|\int_{\mathbb{S}^{d-1}} f(y)\tilde{K}_n(\langle y,z\rangle)d\mu(y)\right|^p d\mu(z) \\ &\le c_3^p C_2^p \|f\|_p^p \le c_3^p C_2^p \|f\|_{W_p^r}^p. \end{aligned}$$

On the other hand, $\sum_{j=1}^m \lambda_j = \int_{\mathbb{S}^{d-1}} d\mu(x) = 1$ by taking $f=1$ in Lemma 2. Then combining the above analysis with (23), we have

$$\begin{aligned} &\left\|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\ \tilde{K}_n(\langle z_j,\cdot\rangle) - c^{(J+2)}\cdot h^{(J+2)}(x) + A\right\|_p \\ &\le \frac{c_5 3^{d+2}c_3^p C_2^p n^{d+3}\|f\|_{W_p^r(\mathbb{S}^{d-1})}}{N^2}. \end{aligned}$$

This together with (21) implies

$$\begin{aligned} &\left\|f - \left(c^{(J+2)}\cdot h^{(J+2)}(x) - A\right)\right\|_{L_p(\mathbb{S}^{d-1})} \\ &\le c_6\|f\|_{W_p^r(\mathbb{S}^{d-1})}\max\left\{n^{-r}, \frac{n^{d+3}}{N^2}\right\}, \end{aligned} \tag{24}$$

where c_6 is a constant depending on d and p given by $c_1 2^{d-1} + c_5 3^{d+2}c_3^p C_2^p$.

The proof for $p=\infty$ is similar:

$$\begin{aligned} &\left\|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\tilde{K}_n(\langle z_j,\cdot\rangle) - c^{(J+2)}\cdot h^{(J+2)}(x) + A\right\|_\infty \\ &= \left\|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\mathcal{B}(\langle z_j,x\rangle)\right\|_\infty \\ &\le \left|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\right| \|\mathcal{B}(\langle z_j,x\rangle\|_\infty. \end{aligned}$$

Using (20) in Lemma 2 when $p=\infty$ and a similar bound as (19) in Lemma 1, we have

$$\left|\sum_{j=1}^m \lambda_j\alpha_n(f)(z_j)\right| \le c_3C_2\|f\|_{W_\infty^r}.$$

Then combining (23) with (21), we get the same bound as (24).

Take $\hat{c}_d = \lfloor c_2+1\rfloor d$ to be a constant depending only on d. When $J \ge \frac{\hat{c}_d}{S-1}$, we take $n = \left\lfloor\left(\frac{(S-1)J}{\lfloor c_2+1\rfloor d}\right)^{\frac{1}{d-1}}\right\rfloor$ and know that $n\in\mathbb{N}$. Take $m = \lfloor c_2+1\rfloor n^{d-1}$, then

$$\begin{aligned} \frac{md}{S-1} &= \frac{\lfloor c_2+1\rfloor n^{d-1}d}{S-1} = \frac{\hat{c}_d}{S-1}n^{d-1} \\ &\le \frac{\hat{c}_d}{S-1}\frac{(S-1)J}{\hat{c}_d} \le J, \end{aligned}$$

so the restriction $J \ge \lceil\frac{md}{S-1}\rceil$ for our construction of CNNs is satisfied. Since $N = \left\lceil\left(\frac{(S-1)J}{\hat{c}_d}\right)^{\frac{d+3+r}{2(d-1)}}\right\rceil$, we know that $N^2 \ge n^{d+3+r}$. Hence

$$\begin{aligned} &\left\|f - \left(c^{(J+2)}\cdot h^{(J+2)}(x) - A\right)\right\|_{L_p(\mathbb{S}^{d-1})} \\ &\le c_6\|f\|_{W_p^r(\mathbb{S}^{d-1})}n^{-r} \\ &\le c_6\|f\|_{W_p^r(\mathbb{S}^{d-1})}2^r\left(\frac{\hat{c}_d}{S-1}\right)^{\frac{r}{d-1}} J^{-\frac{r}{d-1}}. \end{aligned}$$

This verifies the desired error estimate (6) with $C = c_6 2^r\left(\frac{\hat{c}_d}{S-1}\right)^{\frac{r}{d-1}}$.

The total number of free parameters in our network can be bounded as

$$\begin{aligned}\mathcal{N} &\le J(3S+2)+m+2N+4\\ &\le (3S+2)J+\lfloor c_2+1\rfloor\frac{(S-1)}{\hat{c}_d}J\\ &\quad+2\left(\frac{(S-1)}{\hat{c}_d}\right)^{\frac{d+3+r}{2(d-1)}}J^{\frac{d+3+r}{2(d-1)}}+5\\ &\le (3S+5)J^{\max\left\{1,\frac{d+3+r}{2(d-1)}\right\}}+5.\end{aligned}$$

The proof of the theorem is complete. ∎

C. Proof of Corollary 1

Proof of Corollary 1: Let f be a spherical polynomial of degree n. By (17), $f(x)=\int_{\mathbb{S}^{d-1}}f(y)K_n(\langle x,y\rangle)d\sigma(y)$. Since $f(y)\cdot K_n(\langle x,y\rangle)$ is a spherical polynomial in y of degree at most $3n$, by Lemma 2 with $m=\lfloor c_2+1\rfloor n^{d-1}$, there exist $z_1,\ldots,z_m\in\mathbb{S}^{d-1}$ and $\lambda_1,\ldots,\lambda_m>0$ such that

$$f(x)=\sum_{j=1}^{m}\lambda_j f(z_j)K_n(\langle x,z_j\rangle),\quad\forall\ x\in\mathbb{S}^{d-1}.$$

Using the same procedure as in the proof of Theorem 1 with $\alpha_n(f)(z_j)$ simplified by $f(z_j)$, we can construct a CNN network with parameters given explicitly and the final output given as

$$\begin{aligned}&c^{(J+2)}\cdot h^{(J+2)}(x)-A\\ &=\sum_{j=1}^{m}\lambda_j f(z_j)\left[\sum_{\ell=-N-1}^{N+1}\mathcal{A}_\ell(K_n)\sigma\left(\langle z_j,x\rangle-\frac{\ell}{N}\right)\right].\end{aligned}$$

We have

$$\begin{aligned}&\left\|f-c^{(J+2)}\cdot h^{(J+2)}+A\right\|_p\\ &\le\left\|\sum_{j=1}^{m}\lambda_j f(z_j)K_n(\langle z_j,\cdot\rangle)-c^{(J+2)}\cdot h^{(J+2)}+A\right\|_p\\ &\le Cn^{d+3}\|f\|_p/N^2.\end{aligned}$$

The total number of free parameters in the network can be bounded as $\mathcal{N}\le J(3S+2)+n^{d-1}+2N+4$. ∎

IV. Generalization Error Bounds

In this section, we derive learning rates stated in Theorem 2 for the excess misclassification error $\mathcal{R}(\mathrm{sgn}(\widehat{f}_{\mathbf{z}}))-\mathcal{R}(f_c)$ of the classifier $\mathrm{sgn}(\widehat{f}_{\mathbf{z}})$ induced by the CNN network (8). This is achieved by bounds for the excess generalization error $\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)$ together with a comparison theorem below. Here, due to the binary classification nature with the output space $Y=\{-1,1\}$, we can project a real-valued function $f:X\to\mathbb{R}$ onto the interval $[-1,1]$ without changing the sign $\mathrm{sgn}(f)=\mathrm{sgn}(\pi(f))$ by applying the projection operator π defined in [7] as

$$\pi(f)(x):=\begin{cases}1, & \text{if } f(x)>1,\\ -1, & \text{if } f(x)<-1,\\ f(x), & \text{if } -1\le f(x)\le 1.\end{cases}\tag{25}$$

Our generalization error bounds are stated as follows with $\mathscr{N}(\mathcal{H},\eta)$ being the covering numbers of a compact subset $\mathcal{H}$ of $C(X)$ and $\eta>0$ defined as the minimal $l\in\mathbb{N}$ such that there exist $f_1,\ldots,f_\ell\in\mathcal{H}$ satisfying $\mathcal{H}=\bigcup_{i=1}^{\ell}\{g\in\mathcal{H}:\|f_i-g\|_\infty\le\eta\}$.

Theorem 5: Let $\mathcal{H}$ be a compact subset of $C(X)$ with $B=\sup_{f\in\mathcal{H}}\|f\|_\infty$. Define $\widehat{f}_{\mathbf{z}}$ by (8) with a convex loss function $\phi:\mathbb{R}\to\mathbb{R}_+$ and a random sample $\mathbf{z}$. If $\phi(1)=0$ and the pair (ϕ,ρ) has a variancing power $\tau\in[0,1]$ defined by (10), then for any $0<\delta<1$, with the probability at least $1-\delta$, the excess generalization error $\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)$ can be bounded by

$$4\mathcal{D}(\mathcal{H})+\frac{8C_0'\log\frac{2}{\delta}}{3\mathcal{M}}+2\left(\frac{8C_1\log\frac{2}{\delta}}{\mathcal{M}}\right)^{1/(2-\tau)}+24\epsilon^*,\tag{26}$$

where $C_0':=\|\phi\|_{L_\infty[-\max\{B,1\},\max\{B,1\}]}$ and ϵ^* is the smallest positive number ϵ satisfying

$$\mathscr{N}\left(\mathcal{H},\frac{\epsilon}{|\phi_+'(-1)|}\right)\exp\left\{-\frac{\mathcal{M}\epsilon^{2-\tau}}{2C_1+\frac{4}{3}\phi(-1)\epsilon^{1-\tau}}\right\}\le\frac{\delta}{2}.\tag{27}$$

A proof of Theorem 5 can be found in Appendix C.

To apply Theorem 5 to proving Theorem 2, we need bounds for the covering numbers $\mathscr{N}(\mathcal{H},\eta)$, which are given in the following lemma to be proved in Appendix B.

Lemma 4: For integers $S,R,\mathcal{D}_1,\mathcal{D}_2\ge 1$ and $J,N\ge 2$, let $\mathcal{H}:=\mathfrak{H}_{J,\mathcal{D}_1,\mathcal{D}_2,S,R}$ as defined in (7) with $\mathcal{D}_1=(2N+3)\mathcal{D}_2$. For any $\eta>0$, the covering numbers $\mathscr{N}(\mathcal{H},\eta)$ satisfy that

$$\mathscr{N}(\mathcal{H},\eta)\le\left(\frac{40\mathcal{D}_1JR^{J+3}S^J}{\eta}\right)^{J(3S+2)+\mathcal{D}_2+2N+4}.$$

The following comparison theorem established in [7], [36], [3] allows us to estimate the excess misclassification error (9) by the excess generalization error $\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)$ by taking $f=\pi(\widehat{f}_{\mathbf{z}})$.

Lemma 5: Let $f:X\to\mathbb{R}$ be a measurable function. For the p-norm loss $\phi(v)=(1-v)_+^p$ with $p>1$, there holds

$$\mathcal{R}(\mathrm{sgn}(f))-\mathcal{R}(f_c)\le\sqrt{2\left(\mathcal{E}(f)-\mathcal{E}(f_\rho^\phi)\right)}.\tag{28}$$

For the hinge loss $\phi(v)=(1-v)_+$, we have $f_\rho^\phi=f_c$ and

$$\mathcal{R}(\mathrm{sgn}(f))-\mathcal{R}(f_c)\le\mathcal{E}(f)-\mathcal{E}(f_c).\tag{29}$$

Now we are in a position to prove Theorem 2.

Proof of Theorem 2: Denote

$$\mathcal{P}:=J(3S+2)+\mathcal{D}_2+2N+4.$$

By Lemma 4 and the value $|\phi_+'(-1)|=p2^{p-1}$ for the p-norm loss, we know

$$\begin{aligned}&\log\mathscr{N}(\mathcal{H},\frac{\epsilon}{|\phi_+'(-1)|})\\ &\le\mathcal{P}\left(J\log(RS)+3\log\mathcal{P}+\log\left(40p2^{p-1}R^3/\epsilon\right)\right).\end{aligned}$$

Hence for $0<\delta<1$, the quantity ϵ^* defined by (27) can be bounded by any positive solution $\widetilde{\epsilon}$ of the following inequality

$$h(\epsilon)\le\log\frac{\delta}{2}.$$

where $h:\mathbb{R}_+\to\mathbb{R}$ is a decreasing function defined by

$$h(\epsilon)=\mathcal{P}\left(J\log(RS)+3\log\mathcal{P}\right)+\mathcal{P}\log\left(40p2^{p-1}R^3/\epsilon\right)-\frac{\mathcal{M}\epsilon^{2-\tau}}{2C_1+\frac{2^{p+2}}{3}\epsilon^{1-\tau}}.$$

Denote $A:=\mathcal{P}\left(J\log(RS)+3\log\mathcal{P}\right)$. Take

$$\widetilde{\epsilon}=a\left(A\log\frac{2\mathcal{M}}{\delta}/\mathcal{M}\right)^{1/(2-\tau)} \tag{30}$$

with $a:=\max\left\{40p2^{p-1}R^3,C_1\right\}$. Under the restriction

$$A\log\frac{2\mathcal{M}}{\delta}/\mathcal{M}\le 1, \tag{31}$$

we know from the bound $a=\max\left\{40p2^{p-1}R^3,C_1\right\}\ge 40$ that

$$2C_1+\frac{2^{p+2}}{3}\widetilde{\epsilon}^{1-\tau}\le 2C_1+\frac{2^{p+2}}{3}a\le 2^{p+1}a.$$

It follows that

$$\begin{aligned}h(\widetilde{\epsilon})&\le A+\frac{\mathcal{P}}{2-\tau}\log\mathcal{M}-\frac{a^{2-\tau}A\log\frac{2\mathcal{M}}{\delta}}{2^{p+1}a}\\&\le A+\frac{\mathcal{P}}{2-\tau}\log\mathcal{M}-\frac{A\log\frac{2\mathcal{M}}{\delta}}{2^{p+1}}.\end{aligned}$$

If we restrict $\mathcal{M}$ and J further as

$$\mathcal{M}\ge\exp\{2^{p+3}\},\qquad J\ge 2^{p+3}, \tag{32}$$

we see that

$$\begin{aligned}h(\widetilde{\epsilon})&\le\frac{1}{4}\frac{A\log\frac{2\mathcal{M}}{\delta}}{2^{p+1}}+\frac{1}{4}\frac{A\log\frac{2\mathcal{M}}{\delta}}{2^{p+1}}-\frac{A\log\frac{2\mathcal{M}}{\delta}}{2^{p+1}}\\&=-\frac{A\log\frac{2\mathcal{M}}{\delta}}{2^{p+2}}\le-\log\frac{2\mathcal{M}}{\delta}\le\log\frac{\delta}{2},\end{aligned}$$

which implies

$$\epsilon^*\le\widetilde{\epsilon}=a\left(A\log\frac{2\mathcal{M}}{\delta}/\mathcal{M}\right)^{1/(2-\tau)}.$$

Furthermore, since $\hat{c}_d\ge d$ and $N=\left\lceil\left(\frac{(S-1)J}{\hat{c}_d}\right)^{\frac{d+3+r}{2(d-1)}}\right\rceil$, we have

$$\mathcal{P}=J(3S+2)+\mathcal{D}_2+2N+4\le(3S+8)J^\beta$$

and

$$A\le\tilde{C}_4J^{\beta+1}, \tag{33}$$

where $\beta=\max\left\{1,\frac{d+3+r}{2(d-1)}\right\}$ and

$$\tilde{C}_4:=(3S+8)\left(\log(RS)+3\log(3S+8)+3\beta\right).$$

It follows that

$$\epsilon^*\le a\tilde{C}_4^{1/(2-\tau)}\left(\frac{J^{\beta+1}\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{1}{2-\tau}}\left(\log\frac{2}{\delta}\right)^{\frac{1}{2-\tau}}.$$

Putting this bound and condition (11) into Theorem 5, we know that with confidence at least $1-\delta$,

$$\begin{aligned}&\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)\\&\le\tilde{C}_5\left(J^{-\frac{pr}{d-1}}+\left(\frac{J^{\beta+1}\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{1}{2-\tau}}\right)\log\frac{2}{\delta},\end{aligned}$$

where

$$\tilde{C}_5:=4C_0+3C_0'+2\left(8C_1\right)^{1/(2-\tau)}+24a\tilde{C}_4^{1/(2-\tau)}.$$

Thus, we choose

$$J=\left\lceil(\mathcal{M}/\log\mathcal{M})^{\frac{d-1}{(\beta+1)(d-1)+pr(2-\tau)}}\right\rceil$$

and know that with confidence at least $1-\delta$,

$$\begin{aligned}&\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)\\&\le(1+2^{\beta+1})\tilde{C}_5\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{pr}{(\beta+1)(d-1)+pr(2-\tau)}}\log\frac{2}{\delta},\end{aligned} \tag{34}$$

which together with (28) implies

$$\begin{aligned}&\mathcal{R}(\operatorname{sgn}(\widehat{f}_{\mathbf{z}}))-\mathcal{R}(f_c)\\&\le\sqrt{2(1+2^{\beta+1})\tilde{C}_5}\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}}\log\frac{2}{\delta}.\end{aligned} \tag{35}$$

When the restriction (31) is not satisfied, there holds $\log\frac{2\mathcal{M}}{\delta}\ge\frac{\mathcal{M}}{A}$. Then we use the bound (33) and the choice of J to find

$$\log\frac{2\mathcal{M}}{\delta}\ge\frac{\mathcal{M}}{A}\ge\frac{1}{2^{\beta+1}\tilde{C}_4}\mathcal{M}^{\frac{pr(2-\tau)}{(\beta+1)(d-1)+pr(2-\tau)}}. \tag{36}$$

Write $\log\frac{2\mathcal{M}}{\delta}=\log\mathcal{M}+\log\frac{2}{\delta}$. We can choose a constant $\tilde{C}_6>0$ such that

$$\log\mathcal{M}\le\frac{1}{2^{\beta+2}\tilde{C}_4}\mathcal{M}^{\frac{pr(2-\tau)}{(\beta+1)(d-1)+pr(2-\tau)}}$$

whenever $\mathcal{M}\ge\tilde{C}_6$. Combining this with (36), we know that when $\mathcal{M}\ge\tilde{C}_6$, we have

$$\log\frac{2}{\delta}\ge\frac{1}{2^{\beta+2}\tilde{C}_4}\mathcal{M}^{\frac{pr(2-\tau)}{(\beta+1)(d-1)+pr(2-\tau)}},$$

which implies

$$\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}}\log\frac{2}{\delta}\ge\frac{1}{2^{\beta+2}\tilde{C}_4}.$$

But $\mathcal{R}(\operatorname{sgn}(\widehat{f}_{\mathbf{z}}))-\mathcal{R}(f_c)\le 1$, so we also have

$$\begin{aligned}&\mathcal{R}(\operatorname{sgn}(\widehat{f}_{\mathbf{z}}))-\mathcal{R}(f_c)\\&\le 2^{\beta+2}\tilde{C}_4\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}}\log\frac{2}{\delta}.\end{aligned} \tag{37}$$

By the choice of J, we know that there exists an integer $\tilde{C}_7\ge\tilde{C}_6$ such that both restrictions in (32) are satisfied for $\mathcal{M}\ge\tilde{C}_7$. It follows that either (35) with confidence at least $1-\delta$ or (37) is satisfied whenever $\mathcal{M}\ge\tilde{C}_7$.

When $\mathcal{M}<\tilde{C}_7$, we notice

$$\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}}\log\frac{2}{\delta}\ge\tilde{C}_7^{-\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}}.$$

Combining the above analysis, we know that for any $\mathcal{M}\in\mathbb{N}$, with confidence at least $1-\delta$, the desired error bound for Part (i) of Theorem 2 holds true with the constant

$$\widetilde{C}:=\sqrt{2(1+2^{\beta+1})\tilde{C}_5}+2^{\beta+2}\tilde{C}_4+\tilde{C}_7^{\frac{pr}{2(\beta+1)(d-1)+2pr(2-\tau)}}.$$

The bound for the case of the hinge loss with $p=1$ can be verified in the same way by using (29). The proof of the theorem is complete. ■

V. Improved Rates under Noise Conditions

In this section, we prove Theorem 3 which improves the learning rate in Theorem 2 under Tsybakov noise conditions [31]. To this end, we use a Tsybakov function $T = T_\rho : [0,1] \to [0,1]$ motivated by Tsybakov conditions defined in [29] for a probability distribution ρ on $X \times Y$ as

$$T(s) = \rho_X\{x \in X : 0 < |f_\rho(x)| \le s\}.$$

The following result follows easily from arguments in [29] for the least squares loss. We give a proof for completeness.

Proposition 1: Let ϕ be the 2-norm loss function, i.e. $\phi(v) = (1-v)_+^2$ for $v \in \mathbb{R}$. Then $f_\rho^\phi = f_\rho$ and for any measurable function $f : X \to \mathbb{R}$ and $0 < t < 1$,

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \le T\left(\sqrt{\left(\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho^\phi)\right)/t}\right) + t. \tag{38}$$

Proof: For any measurable function $g : X \to [-1,1]$, we have $\phi(yg(x)) = (1-yg(x))^2 = (y^2 - yg(x))^2 = (y-g(x))^2$ which equals the least squares loss. Hence $f_\rho^\phi = f_\rho : X \to [-1,1]$.

For a measurable function $f : X \to \mathbb{R}$, we denote

$$X_f := \{x \in X : \mathrm{sgn}(f(x)) \ne \mathrm{sgn}(f_\rho(x)), |f_\rho(x)| > 0\}.$$

It is well known (see [21] or [29]) that

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) = \int_{X_f} |f_\rho(x)| d\rho_X.$$

We can also see from the proof of Proposition 2 in [29] that

$$\rho_X(X_f) \le T\left(\|f - f_\rho\|_\infty\right),$$

and for any $t > 0$,

$$\rho_X(X_f) \le T\left(\|f - f_\rho\|_{L^2_{\rho_X}}/\sqrt{t}\right) + t.$$

But the projected function $\pi(f) : X \to [-1,1]$ satisfies $X_{\pi(f)} = X_f$ and

$$\|\pi(f) - f_\rho\|^2_{L^2_{\rho_X}} = \int_Z (y - \pi(f)(x))^2 d\rho - \int_Z (y - f_\rho(x))^2 d\rho$$

which equals $\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho^\phi)$, so the above estimate yields

$$\rho_X(X_f) \le T\left(\sqrt{\left(\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho^\phi)\right)/t}\right) + t.$$

Since $|f_\rho(x)| \le 1$, our desired bound follows. ∎

Proof of Theorem 3: Tsybakov noise condition (14) yields $T(s) = O(s^\theta)$. Taking $t = \Delta^{\theta/(2+\theta)}$ for $\Delta > 0$ gives

$$T\left(\sqrt{\Delta/t}\right) + t = O\left(\Delta^{\theta/(2+\theta)}\right).$$

This together with (38) in Proposition 1 applied to $f = \widehat{f_{\mathbf{z}}}$ tells us that

$$\mathcal{R}(\mathrm{sgn}(\widehat{f_{\mathbf{z}}})) - \mathcal{R}(f_c) = O\left(\mathcal{E}(\pi(\widehat{f_{\mathbf{z}}})) - \mathcal{E}(f_\rho^\phi))^{\frac{\theta}{2+\theta}}\right). \tag{39}$$

The 2-norm loss has a variancing power $\tau = 1$. So we combine (39) with (34) for $p = 2$ and $\tau = 1$ to conclude that with confidence at least $1-\delta$,

$$\mathcal{R}(\mathrm{sgn}(\widehat{f_{\mathbf{z}}})) - \mathcal{R}(f_c) \le \widetilde{C}\left(\frac{\log \mathcal{M}}{\mathcal{M}}\right)^{\frac{2r\theta}{(2+\theta)((\beta+1)(d-1)+2r)}} \log\frac{2}{\delta}.$$

This proves Theorem 3. ∎

Proof of Theorem 4: By Corollary 1 and the same procedure as in the proof of Theorem 2, Theorem 5 implies that with confidence $1-\delta$,

$$\mathcal{E}(\pi(\hat{f}_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le C\left[\left(\frac{N\log\mathcal{M}}{\mathcal{M}}\right) + N^{-4}\right]\log\frac{2}{\delta}.$$

Taking $N = \left(\frac{\mathcal{M}}{\log\mathcal{M}}\right)^{1/5}$ implies that

$$\mathcal{E}(\pi(\hat{f}_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le C\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{-4/5}\log\frac{2}{\delta}.$$

Now applying Proposition 1 and Tsybakov condition (14),

$$\mathcal{R}(\mathrm{sgn}(\hat{f}_{\mathbf{z}})) - \mathcal{R}(f_c) \le C\left[\left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{2\theta/5} t^{-\theta/2} + t\right]\log\frac{2}{\delta},$$

which completes the proof by taking $t = \left(\frac{\log\mathcal{M}}{\mathcal{M}}\right)^{4\theta/5(2+\theta)}$. ∎

VI. Numerical Experiments

In this section, we illustrate our theoretical results by numerical experiments. We study the misclassification error with the least squares loss and hinge loss.

We consider the following probability model for generating simulated data. Let X be a random vector uniformly distributed in $\mathbb{S}^{d-1}$. For given $X = \mathbf{x} = (x_1, x_2, \ldots, x_d)$, $Y \in \{1, -1\}$ follows a Bernoulli distribution $\Pr(Y = 1 \mid \mathbf{x}) = h(\mathbf{x})$. We take $h(\mathbf{x}) = \frac{2f(\mathbf{x}) - f_{\max} - f_{\min}}{f_{\max} - f_{\min}}$, where

$$f(\mathbf{x}) = \sum_{i=1}^{3} \varphi(\|\mathbf{x} - e_i\|/1.5\pi), \text{ for } \mathbf{x} \in \mathbb{S}^{d-1},$$

with $\varphi(r) = (1-r)_+^4(4r+1)$, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, and $f_{\max}, f_{\min}$ are maximum and minimum values of f respectively.

We estimate the classifier by using the least squares loss as well as the hinge loss with the CNN architecture of $0.5\mathcal{M}^{1/3}$ layers followed by two fully connected layers having $50, 35$ neurons, respectively. Maxpooling with kernel size 10 is taken after the convolutional layers. In our experiment, we take various training data sizes $\mathcal{M} = \{50, 100, 300, 1000, 2000, 5000, 10000\}$ and two different input dimensions of $10, 50$. For evaluation, we randomly simulate 2000 sample points from a test set with Bayes error 3.6% for dim10 and 0.85% for dim50. For training the proposed CNN, as in the literature [22], we use Adam with epoch 5, learning rate 10^{-2} and mini batch size $\mathcal{M}/20$.

Table III presents the classification accuracy of the CNN classifier over 10 independent trails, and Fig. 1, Fig. 2 draw the trace plots of the excess misclassification error for various sample sizes. The results confirm well the theoretical results that the excess misclassification error converges to 0 as the sample size increases.

TABLE III
CLASSIFICATION ACCURACY

dim=10							
	50	**100**	**300**	**1000**	**2000**	**5000**	**10000**
MSE	63.75	83.80	90.85	94.55	95.35	95.85	95.09
Hinge	44.50	77.80	81.19	93.85	93.95	95.85	95.34
dim=50							
	50	**100**	**300**	**1000**	**2000**	**5000**	**10000**
MSE	75.85	91.09	96.45	96.19	95.60	96.80	97.80
Hinge	51.95	88.30	89.59	90.25	96.19	95.69	96.90

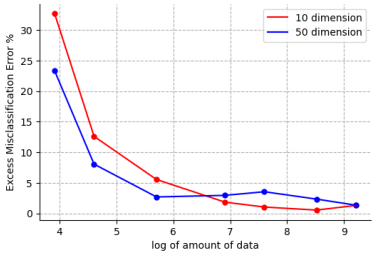


Fig. 1. Excess misclassification error with respect to different sample sizes and the least squares loss

VII. CONCLUSION

In this paper we have studied the L_p approximation by deep CNNs with $1 \leq p \leq \infty$ using efficient cubature formulae and some other techniques from spherical analysis and approximation theory. Based on the obtained approximation orders, we have established excess misclassification error rates of deep CNN-based classifiers by deriving some generalization error bounds with respect to the $p-$norm loss. In addition, we showed that CNN-based classifiers can achieve the almost optimal rate by imposing a Tsybakov noise condition on the data distribution.

APPENDIX A
STRUCTURE OF DEEP NETWORK FOR APPROXIMATION

The main idea to approximate the discretization of the integral linear operator $L_n(f)$, i.e. $\sum_{j=1}^m \lambda_j \alpha_n(f)(z_j)\tilde{K}_n(\langle z_j, \cdot\rangle)$ is conducted by approximating ridge functions $\tilde{K}_n(\langle z_j, \cdot\rangle)$ using deep CNNs. In this part, we describe the structure of our construction.

We first use J layers to realize features $\{\langle z_j, x\rangle\}_{j=1}^m$, where we apply the following two lemmas proved in [37] by stacking $\{z_j \in \mathbb{S}^{d-1}\}_{j=1}^m$ into a sequence W and factorizing it into convolutions of filters.

Lemma 6: Let $S \geq 2$ and $W = (W_k)_{k=-\infty}^{\infty}$ be a sequence supported in $\{0, \cdots, \mathcal{L}\}$ with $\mathcal{L} \geq 0$. Then there exists a finite sequence of filters $\left\{w^{(j)}\right\}_{j=1}^p$ each supported in $\{0, \cdots, S\}$ with $p \leq \lceil \frac{\mathcal{L}}{S-1} \rceil$ such that the following convolutional factorization holds true

$$W = w^{(p)} * w^{(p-1)} * \cdots * w^{(2)} * w^{(1)}.$$

Lemma 7: Let $\{w^{(k)}\}_{k=1}^J$ be a set of sequences supported in $\{0, 1, \ldots, S\}$. Then

$$T^{(J)} \cdots T^{(2)} T^{(1)} = T^{(J,1)} := (W_{i-k})_{i=1,\ldots,d+JS, k=1,\ldots,d} \tag{40}$$

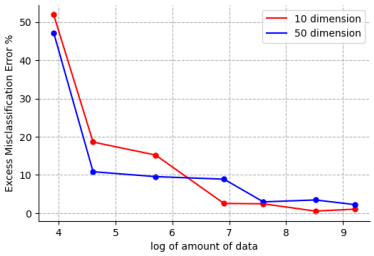


Fig. 2. Excess misclassification error with respect to different sample sizes and the hinge Loss

is a Toeplitz matrix in $\mathbb{R}^{(d+JS)\times d}$ associated with the filter $W = w^{(J)} * \cdots * w^{(2)} * w^{(1)}$ supported in $\{0, 1, \cdots, JS\}$. Define a sequence W supported in $\{0, 1, \ldots, md-1\}$ by stacking $\{z_j\}_{j=1}^m$ in reversed orders as $W_{(j-1)d+(d-i)} = (z_j)_i$, $i \in \{1, \ldots, d\}$. It is constructed in such a way that the (jd)-th row of the matrix $T^{(J,1)}$ is exactly the row vector z_j^T. From Lemma 6, we can get a sequence of filters $\left\{w^{(j)}\right\}_{j=1}^J$ supported in $\{0, 1, \ldots, S\}$ with $J \geq \lceil \frac{md}{S-1} \rceil$ satisfying $W = w^{(J)} * w^{(J-1)} * \cdots * w^{(2)} * w^{(1)}$. Note that here we take $\{w^{(j)}\}_{j=p+1}^J$ to be the delta sequence δ_0 given by $(\delta_0)_0 = 1$ and $(\delta_0)_k = 0$ for $k \in \mathbb{Z} \setminus \{0\}$. Then from Lemma 7, these filters yield a Toeplitz type convolutional matrix belonging to $\mathbb{R}^{(d+JS)\times d}$, that is,

$$T^{(J)} \cdots T^{(2)} T^{(1)} = T^{(J,1)} = (W_{i-k})_{i=1,\ldots,d+JS, k=1,\ldots,d}.$$

Then we construct the bias vectors $b^{(j)}$ satisfying $b_{S+1}^{(j)} = \cdots = b_{d_j-S}^{(j)}$ for $j = 1, \ldots, J$. Let $b^{(1)} = -\left\|w^{(1)}\right\|_1 \mathbf{1}_{d_0}$ and

$$\begin{aligned} b^{(j)} = &\left(\Pi_{p=1}^{j-1} \left\|w^{(p)}\right\|_1\right) T^{(j)} \mathbf{1}_{d_{j-1}} \\ &- \left(\Pi_{p=1}^{j} \left\|w^{(p)}\right\|_1\right) \mathbf{1}_{d_{j-1}+S} \end{aligned} \tag{41}$$

for $j = 2, \ldots, J$, where $\|w\|_1 = \sum_{k=-\infty}^{\infty} |w_k|$ and $\mathbf{1}_\ell$ is the constant $\mathbf{1}$ vector in $\mathbb{R}^\ell$. With the choice of these bias vectors we know from [38, Lemma 3] that

$$h^{(J)}(x) = T^{(J)} \cdots T^{(2)} T^{(1)} x + B^{(J)} \mathbf{1}_{d_J} = T^{(J,1)} x + B^{(J)} \mathbf{1}_{d_J},$$

where $B^{(J)} = \Pi_{p=1}^J \left\|w^{(p)}\right\|_1$. Together with the definition of W we know the components of $h^{(J)}(x)$ satisfy

$$\left(h^{(J)}(x)\right)_{jd} = \langle z_j, x\rangle + B^{(J)}, \qquad j = 1, \ldots, m.$$

After applying the downsampling operator to $h^{(J)}$ we get the features, that is,

$$\mathfrak{D}_d\left(h^{(J)}(x)\right) = \begin{bmatrix} \langle z_1, x\rangle \\ \vdots \\ \langle z_m, x\rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix} + B^{(J)} \mathbf{1}_{\lfloor (d+JS)/d \rfloor} \in \mathbb{R}^{\mathcal{D}_2}. \tag{42}$$

The next step is to construct a fully connected $(J+1)$-th layer of width $\mathcal{D}_1 = (2N+3)\mathcal{D}_2$ to generate functions $\sigma(\langle z_j, x\rangle - t_i)$ with $t_i = -1 + \frac{i-2}{N}$ for $j = 1, \ldots, m$ and

$i = 1, \ldots, 2N+3$. Here we choose the connection matrix $F^{(J+1)}$ in a block form as

$$F^{(J+1)} = \begin{bmatrix} \mathbf{1}_{2N+3} & O & O\cdots & O \\ O & \mathbf{1}_{2N+3} & O\cdots & O \\ \vdots & \ddots & \ddots & \vdots \\ O & \cdots & O & \mathbf{1}_{2N+3} \end{bmatrix} \in \mathbb{R}^{\mathcal{D}_1 \times \mathcal{D}_2} \tag{43}$$

and the bias vector

$$b^{(J+1)}_{(j-1)(2N+3)+i} = \begin{cases} B^{(J)} + t_i, & \text{if } 1 \le j \le m, \\ B^{(J)} + 1, & \text{if } j > m. \end{cases}$$

Then the first fully connected layer $h^{(J+1)}(x) \in \mathbb{R}^{\mathcal{D}_1}$ of the deep network is

$$\begin{aligned} &\left(h^{(J+1)}\right)_{(j-1)(2N+3)+i} \\ &= \begin{cases} \sigma\left(\langle z_j, \cdot\rangle - t_i\right), & \text{if } j \le m,\ 1 \le i \le 2N+3, \\ 0, & \text{if } j > m. \end{cases} \end{aligned}$$

Note that $h^{(J+1)}$ can be seen as a vector contains $\mathcal{D}_2$ blocks of equal size $2N+3$, and each j-th block represents $\{\sigma\left(\langle z_j, \cdot\rangle - t_i\right)\}_{i=1}^{2N+3}$ for $j = 1, \ldots, m$, while other blocks are zero vectors.

The last step is to use another fully connected layer to produce $Q_N(\tilde{K}_n)$. Notice that

$$Q_N(\tilde{K}_n) = N \sum_{i=1}^{2N+3} \left(\mathcal{L}_N\left(\left\{\tilde{K}_n(t_k)\right\}_{k=2}^{2N+2}\right)\right)_i \sigma\left(\cdot - t_i\right)$$

with $\mathcal{L}_N : \mathbb{R}^{2N+1} \to \mathbb{R}^{2N+3}$ given for $\zeta = (\zeta_i)_{i=1}^{2N+1} \in \mathbb{R}^{2N+1}$ by

$$\left(\mathcal{L}_N(\zeta)\right)_i = \begin{cases} \zeta_2, & \text{for } i = 1, \\ \zeta_3 - 2\zeta_2, & \text{for } i = 2, \\ \zeta_{i-1} - 2\zeta_i + \zeta_{i+1}, & \text{for } 3 \le i \le 2N+1, \\ \zeta_{2N+1} - 2\zeta_{2N+2}, & \text{for } i = 2N+2, \\ \zeta_{2N+2}, & \text{for } i = 2N+3. \end{cases}$$

Take the vector $\Theta_N \in \mathbb{R}^{2N+3}$ in terms of the linear operator $\mathcal{L}_N$ as

$$\Theta_N = N\mathcal{L}_N\left(\left\{\tilde{K}_n(t_i)\right\}_{i=2}^{2N+2}\right).$$

For the $(J+2)$-th fully connected layer, we choose the connection matrix $F^{(J+2)}$ as

$$F^{(J+2)} = \begin{bmatrix} \Theta_N^T & O & O\cdots & O \\ O & \Theta_N^T & O\cdots & O \\ \vdots & \ddots & \ddots & \vdots \\ O & \cdots & O & \Theta_N^T \end{bmatrix} \in \mathbb{R}^{\mathcal{D}_2 \times (2N+3)\mathcal{D}_2}.$$

Observe that the jth component of $F^{(J+2)}h^{(J+1)}(x)$ is the product of Θ_N^T and the jth block of $h^{(J+1)}(x)$. And by taking $B^{(J+2)} = \|\tilde{K}_n\|_{C[-1,1]}$ and

$$b^{(J+2)} = \begin{bmatrix} -B^{(J+2)}\mathbf{1}_m \\ O \end{bmatrix} \in \mathbb{R}^{\mathcal{D}_2 \times (2N+3)\mathcal{D}_2},$$

we know that the last layer $h^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}$ is given by

$$h^{(J+2)}(x) = \begin{bmatrix} \left[Q_N(\tilde{K}_n)\left(\langle z_j, x\rangle\right) + B^{(J+2)}\right]_{j=1}^m \\ O \end{bmatrix}. \tag{44}$$

Finally by taking $c^{(J+2)} \in \mathbb{R}^{\mathcal{D}_2}$ as

$$c_j^{(J+2)} = \begin{cases} \lambda_j \alpha_n(f)(z_j), & \text{if } j = 1, \ldots, m, \\ 0, & \text{otherwise,} \end{cases}$$

and $A = B^{(J+2)} \sum_{j=1}^m \lambda_j \alpha_n(f)(z_j)$, the final output function of our network is

$$c^{(J+2)} \cdot h^{(J+2)}(x) - A = \sum_{j=1}^m \lambda_j \alpha_n(f)(z_j) Q_N(\tilde{K}_n)\left(\langle z_j, x\rangle\right). \tag{45}$$

Appendix B
Bounding the covering numbers

In this part, we use the construction described in the previous appendix to prove Lemma 4.

Proof of Lemma 4: For simplicity, we set

$$C_{S,R} := (S+1)R.$$

Since all filters are bounded as $\|w^{(j)}\|_\infty \le R$, $j = 1, \ldots, J$, the convolutional matrix $T^{(j)}$ associated with $w^{(j)}$ defined by (2) satisfy

$$\|T^{(j)}u\|_\infty \le C_{S,R}\|u\|_\infty, \qquad \forall u \in \mathbb{R}^{d_{j-1}}.$$

Note that $R \ge 1$. We claim that

$$\|h^{(j)}\|_\infty \le 2C_{S,R}^j, \quad j = 0, 1, \ldots, J. \tag{46}$$

In fact, $h^{(0)}(x) = x \in \mathbb{S}^{d-1}$ satisfies $\|h^{(0)}(x)\|_\infty \le 1$. For $j = 1, \ldots, J$, by (1) and the assumptions $\|b^{(j)}\|_\infty \le R$,

$$\begin{aligned} \|h^{(j)}(x)\|_\infty \le& \|T^{(j)}h^{(j-1)}\|_\infty + \|b^{(j)}\|_\infty \\ \le& C_{S,R}\|h^{(j-1)}\|_\infty + R. \end{aligned}$$

By iteration, this yields

$$\begin{aligned} \|h^{(j)}(x)\|_\infty &\le C_{S,R}^j + (1 + C_{S,R} + \cdots + C_{S,R}^{j-1})R \\ &\le C_{S,R}^j + \frac{C_{S,R}^j - 1}{S} \le 2C_{S,R}^j \end{aligned}$$

and verifies our claim.

Consider two functions $f, \tilde{f} \in \mathcal{H}$ generated by two different choices of parameters $\{w^{(j)}, b^{(j)}, F^{(J+1)}, F^{(J+2)}, c, A\}$ and $\{\tilde{w}^{(j)}, \tilde{b}^{(j)}, \tilde{F}^{(J+1)}, \tilde{F}^{(J+2)}, \tilde{c}, \tilde{A}\}$ respectively. For any $\eta > 0$, if $\|w^{(j)} - \tilde{w}^{(j)}\|_\infty \le \eta$, $\|b^{(j)} - \tilde{b}^{(j)}\|_\infty \le \eta$, $\|F^{(J+2)} - \tilde{F}^{(J+2)}\|_\infty \le \eta$, $\|c - \tilde{c}\|_\infty \le \eta$, $\|A - \tilde{A}\|_\infty \le \eta$ and $F^{(J+1)}$ is the same as $\tilde{F}^{(J+1)}$ given by (43), then we claim that the generated layers $\{h^{(j)}, \tilde{h}^{(j)}\}_{j=0}^{J+2}$ satisfy

$$\|h^{(j)} - \tilde{h}^{(j)}\|_\infty \le (2j+2)C_{S,R}^j \eta. \tag{47}$$

The case $j = 0$ is trivial because $h^{(0)}(x) = \tilde{h}^{(0)}(x) = x \in \mathbb{S}^{d-1}$. For $j = 1, \ldots, J$, since $\sigma(u) \le |u|$ for any $u \in \mathbb{R}$, we

have that

$$\begin{aligned}
&\|h^{(j)}-\tilde{h}^{(j)}\|_\infty\\
&\le \|\tilde{T}^{(j)}(h^{(j-1)}-\tilde{h}^{(j-1)})\|_\infty\\
&\quad+\|(T^{(j)}-\tilde{T}^{(j)})h^{(j-1)}\|_\infty+\|b^{(j)}-\tilde{b}^{(j)}\|_\infty\\
&\le C_{S,R}\|h^{(j-1)}-\tilde{h}^{(j-1)}\|_\infty+(S+1)\eta\|h^{(j-1)}\|_\infty+\eta\\
&\le C_{S,R}\|h^{(j-1)}-\tilde{h}^{(j-1)}\|_\infty+2C_{S,R}^j\eta+\eta,
\end{aligned}$$

which verifies our claim

$$\begin{aligned}
\|h^{(j)}-\tilde{h}^{(j)}\|_\infty \le& 2jC_{S,R}^j\eta+[1+C_{S,R}+\cdots+C_{S,R}^j]\eta\\
\le&(2j+2)C_{S,R}^j\eta
\end{aligned}$$

since $C_{S,R}\ge 2$ and $\frac{C_{S,R}^{j+1}-1}{C_{S,R}-1}\le 2C_{S,R}^j$. Note that $F^{(J+1)}=\tilde{F}^{(J+1)}$ and for $v=(v_1,\ldots,v_{\mathcal{D}_2})\in\mathbb{R}^{\mathcal{D}_2}$, $i=0,1,\ldots,\mathcal{D}_2$, $j=1,\ldots,2N+3$,

$$(F^{(J+1)}v)_{(2N+3)i+j}=v_j.$$

Then we have

$$\begin{aligned}
\|h^{(J+1)}\|_\infty &\le \|F^{(J+1)}h^{(J)}\|_\infty+\|b^{(J+1)}\|_\infty\\
&\le \|h^{(J)}\|_\infty+R\le 3C_{S,R}^J
\end{aligned}$$

and

$$\begin{aligned}
&\|h^{(J+1)}-\tilde{h}^{(J+1)}\|_\infty\\
&\le \|(F^{(J+1)}(h^{(J)}-\tilde{h}^{(J)})\|_\infty+\|b^{(J+1)}-\tilde{b}^{(J+1)}\|_\infty\\
&\le \|h^{(J)}-\tilde{h}^{(J)}\|_\infty+\eta\le(2J+3)C_{S,R}^J\eta.
\end{aligned}$$

For the last layer, since $F^{(J+2)},\tilde{F}^{(J+2)}\in\mathbb{R}^{\mathcal{D}_2\times\mathcal{D}_1}$ and each row has only $(2N+3)$ nonzero entries, we have

$$\begin{aligned}
\|h^{(J+2)}\|_\infty &\le \|F^{(J+2)}h^{(J+1)}\|_\infty+\|b^{(J+2)}\|_\infty\\
&\le 3(2N+3)RC_{S,R}^J+R\le 3(2N+4)RC_{S,R}^J
\end{aligned}$$

and

$$\begin{aligned}
&\|h^{(J+2)}-\tilde{h}^{(J+2)}\|_\infty\\
&\le \|(F^{(J+2)}h^{(J+1)}-b^{(J+2)})-(\tilde{F}^{(J+2)}\tilde{h}^{(J+1)}-\tilde{b}^{(J+2)})\|_\infty\\
&\le \|\tilde{F}^{(J+2)}(h^{(J+1)}-\tilde{h}^{(J+1)})\|_\infty+\|b^{(J+2)}-\tilde{b}^{(J+2)}\|_\infty\\
&\quad+\|(F^{(J+2)}-\tilde{F}^{(J+2)})h^{(J+1)}\|_\infty\\
&\le (2N+3)R(2J+3)C_{S,R}^J\eta+3(2N+3)RC_{S,R}^J\eta+\eta\\
&\le (2J+7)(2N+3)RC_{S,R}^J\eta\le 6J(2N+3)RC_{S,R}^J\eta.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\|f-\tilde{f}\|_\infty\\
&\le \|(c-\tilde{c})\cdot h^{(J+2)}\|_\infty+\|\tilde{c}\cdot(h^{(J+2)}-\tilde{h}^{(J+2)})\|_\infty\\
&\quad+\|A-\tilde{A}\|_\infty\\
&\le 3R(2N+4)\mathcal{D}_2RC_{S,R}^J\eta+6J(2N+3)RC_{S,R}^J\mathcal{D}_2R\eta+\eta\\
&\le (18N+31)\mathcal{D}_2JR^2C_{S,R}^J\eta\le 20\mathcal{D}_1JS^JR^{J+2}\eta.
\end{aligned}$$

This implies that an η-net of parameters

$$\left\{w^{(j)},b^{(j)},F^{(J+1)},F^{(J+2)},c,A\right\}\in\mathbb{R}^{J(3S+2)+\mathcal{D}_2+2N+4}$$

yields a $20\mathcal{D}_1JS^JR^{J+2}\eta$-net of $\mathcal{H}$. Therefore, for any $\eta>0$,

$$\mathscr{N}(\mathcal{H},\eta)\le\left(\frac{40\mathcal{D}_1JS^JR^{J+3}}{\eta}\right)^{J(3S+2)+\mathcal{D}_2+2N+4}.$$

This proves the lemma. ∎

APPENDIX C
PROOF OF THEOREM 5

Our basic tools to prove Theorem 5 are the classical Bernstein inequality (Lemma 8) and the following concentration inequality (Lemma 9) which can be found in [11] and [7].

Lemma 8: Let ξ be a random variable on a compact metric space Z with mean $\mathbb{E}(\xi)$ and variance $\sigma^2(\xi)=\sigma^2$ and $\{z_i\}_{i=1}^{\mathcal{M}}$ is an independent random sample. If $|\xi(z)-\mathbb{E}(\xi)|\le b$ for some $b>0$ almost surely, then for any $\epsilon>0$,

$$\text{Prob}\left\{\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\xi(z_i)-\mathbb{E}(\xi)>\epsilon\right\}\le\exp\left\{-\frac{\mathcal{M}\epsilon^2}{2\sigma^2+\frac{2}{3}b\epsilon)}\right\}. \tag{48}$$

Lemma 9: Let $0\le\tau\le 1$, $b>0$, $c>0$, and $\mathcal{G}$ be a function set defined on a probability space (Z,ρ) such that for each $g\in\mathcal{G}$, $\mathbb{E}(g):=\int_Z g(z)d\rho\ge 0$, $\|g-\mathbb{E}(g)\|_\infty\le b$ almost surely and $\mathbb{E}(g^2)\le c(\mathbb{E}(g))^\tau$. Then for any $\epsilon>0$ and random sample $\{z_i\}_{i=1}^{\mathcal{M}}$, there holds

$$\begin{aligned}
&\text{Prob}\left\{\sup_{g\in\mathcal{G}}\frac{\mathbb{E}(g)-\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}g(z_i)}{\sqrt{(\mathbb{E}(g))^\tau+\epsilon^\tau}}>4\epsilon^{1-\tau/2}\right\}\\
&\le\mathscr{N}(\mathcal{G},\epsilon)\exp\left\{-\frac{\mathcal{M}\epsilon^{2-\tau}}{2(c+\frac{1}{3}b\epsilon^{1-\tau})}\right\}.
\end{aligned} \tag{49}$$

Now we are in a position to prove Theorem 5.

Proof of Theorem 5: Since $\phi(1)=0$, by the convexity, we know that ϕ is nondecreasing on $[1,\infty)$ and nonincreasing on $(-\infty,1]$. Hence $f_\rho^\phi(x)\in[-1,1]$ on X and for any $f\in\mathcal{H}$,

$$\mathcal{E}_{\mathbf{z}}(\pi(f))\le\mathcal{E}_{\mathbf{z}}(f),\quad\text{and}\quad\mathcal{E}_{\mathbf{z}}(\widehat{f_{\mathbf{z}}})\le\mathcal{E}_{\mathbf{z}}(f),$$

which implies that

$$\mathcal{E}_{\mathbf{z}}(\pi(\widehat{f_{\mathbf{z}}}))-\mathcal{E}_{\mathbf{z}}(f)\le 0,\quad\forall\ f\in\mathcal{H}.$$

It follows that with $\hat{f}=\arg\inf_{f\in\mathcal{H}}\mathcal{E}(f)$, we have

$$\begin{aligned}
&\mathcal{E}(\pi(\widehat{f_{\mathbf{z}}}))-\mathcal{E}(f_\rho^\phi)\\
&=\{\mathcal{E}(\hat{f})-\mathcal{E}(f_\rho^\phi)\}+\{\mathcal{E}(\pi(\widehat{f_{\mathbf{z}}}))-\mathcal{E}_{\mathbf{z}}(\pi(\widehat{f_{\mathbf{z}}}))\}\\
&\quad+\{\mathcal{E}_{\mathbf{z}}(\pi(\widehat{f_{\mathbf{z}}}))-\mathcal{E}_{\mathbf{z}}(\hat{f})\}+\{\mathcal{E}_{\mathbf{z}}(\hat{f})-\mathcal{E}(\hat{f})\}\\
&\le\{\mathcal{E}(\hat{f})-\mathcal{E}(f_\rho^\phi)\}+\{\mathcal{E}_{\mathbf{z}}(\hat{f})-\mathcal{E}_{\mathbf{z}}(f_\rho^\phi)+\mathcal{E}(f_\rho^\phi)-\mathcal{E}(\hat{f})\}\\
&\quad+\{\mathcal{E}(\pi(\widehat{f_{\mathbf{z}}}))-\mathcal{E}(f_\rho^\phi)+\mathcal{E}_{\mathbf{z}}(f_\rho^\phi)-\mathcal{E}_{\mathbf{z}}(\pi(\widehat{f_{\mathbf{z}}}))\}\\
&=:\mathcal{D}(\mathcal{H})+S_2(\mathcal{H})+S_1(\mathcal{H}).
\end{aligned} \tag{50}$$

We then proceed in three steps.

Step 1: The estimation of $\mathcal{S}_2(\mathcal{H})$.

Consider the random variable ξ on (Z,ρ) given by

$$\xi(z)=\phi(y\hat{f}(x))-\phi(yf_\rho^\phi(x)).$$

Note that $\mathbb{E}(\xi)=\mathcal{E}(\hat{f})-\mathcal{E}(f_\rho^\phi)=\mathcal{D}(\mathcal{H})\ge 0$. Since $\|\hat{f}\|_\infty\le B$, we know that $|\xi(z)|\le C_0'=\|\phi\|_{L_\infty[-\max\{B,1\},\max\{B,1\}]}$ implying $|\xi-\mathbb{E}(\xi)|\le 2C_0'$ almost surely. Since (ϕ,ρ) has a variancing power $\tau\in[0,1]$, we have $\sigma^2(\xi)\le\mathbb{E}(\xi^2)\le C_1\left(\mathcal{D}(\mathcal{H})\right)^\tau$. Therefore, the following holds according to Lemma 8,

$$\begin{aligned}
&\text{Prob}\left\{\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\xi(z_i)-\mathbb{E}(\xi)>\epsilon\right\}\\
&\le\exp\left\{-\frac{\mathcal{M}\epsilon^2}{2C_1\left(\mathcal{D}(\mathcal{H})\right)^\tau+\frac{4}{3}C_0'\epsilon}\right\}:=\frac{\delta}{2}.
\end{aligned}$$

Setting the probability bound to be $\frac{\delta}{2}$ and solving the quadratic equation for ϵ tells us that with confidence at least $1-\delta/2$,

$$\begin{aligned}&\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\xi(z_i)-\mathbb{E}(\xi)\\&\le\frac{\frac{2}{3}C_0'\log\frac{2}{\delta}+\sqrt{\frac{4}{9}(C_0')^2(\log\frac{2}{\delta})^2+8\mathcal{M}C_1\log\frac{2}{\delta}\left(\mathcal{D}(\mathcal{H})\right)^\tau}}{\mathcal{M}}\\&\le\frac{4C_0'\log\frac{2}{\delta}}{3\mathcal{M}}+\sqrt{\frac{8C_1\log\frac{2}{\delta}\left(\mathcal{D}(\mathcal{H})\right)^\tau}{\mathcal{M}}}.\end{aligned}$$

Applying the elementary inequality with the dual number $p'=\frac{p}{p-1}$ of $p\in(1,\infty)$,

$$ab\le\frac{1}{p}a^p+\frac{1}{p'}b^{p'},\quad\forall a,b>0 \tag{51}$$

to $p=\frac{2}{2-\tau}$, $p'=\frac{2}{\tau}$, $a=\sqrt{8C_1\log\frac{2}{\delta}/\mathcal{M}}$ and $b=\sqrt{\left(\mathcal{D}(\mathcal{H})\right)^\tau}$, we get

$$\begin{aligned}&\sqrt{\frac{8C_1\log\frac{2}{\delta}\left(\mathcal{D}(\mathcal{H})\right)^\tau}{\mathcal{M}}}\\&\le(1-\frac{\tau}{2})\left(\frac{8C_1\log\frac{2}{\delta}}{\mathcal{M}}\right)^{1/(2-\tau)}+\frac{\tau}{2}\mathcal{D}(\mathcal{H}).\end{aligned}$$

Hence with confidence at least $1-\delta/2$, we have

$$\begin{aligned}&\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\xi(z_i)-\mathbb{E}(\xi)\\&\le\frac{4C_0'\log\frac{2}{\delta}}{3\mathcal{M}}+\left(\frac{8C_1\log\frac{2}{\delta}}{\mathcal{M}}\right)^{1/(2-\tau)}+\mathcal{D}(\mathcal{H}).\end{aligned} \tag{52}$$

Step 2: The estimation of $\mathcal{S}_1(\mathcal{H})$.

Choose a function set $\mathcal{G}=\{g(z)\phi(y\pi(f)(x))-\phi(yf_\rho^\phi(x)):f\in\mathcal{H}\}$, then for every $g\in\mathcal{G}$, $\|g\|_\infty\le\phi(-1)$, $\mathbb{E}(g)=\mathcal{E}(\pi(f))-\mathcal{E}(f_\rho^\phi)\ge0$, $\|g-\mathbb{E}(g)\|_\infty\le2\phi(-1)$ and $\mathbb{E}(g^2)\le C_1(\mathbb{E}(g))^\tau$. Note that for every $g_1,g_2\in\mathcal{G}$ we know $\|g_1-g_2\|_\infty\le|\phi_+'(-1)|\|f_1-f_2\|_\infty$, which means an $\frac{\epsilon}{|\phi_+'(-1)|}$-cover of $\mathcal{H}$ generates an ϵ-cover of $\mathcal{G}$, that is

$$\mathscr{N}(\mathcal{G},\epsilon)\le\mathscr{N}\left(\mathcal{H},\frac{\epsilon}{|\phi_+'(-1)|}\right).$$

Applying Lemma 9 to $\mathcal{G}$, we obtain, for any $\epsilon>0$ and $\forall g\in\mathcal{G}$,

$$\mathbb{E}(g)-\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}g(z_i)\le4\epsilon^{1-\tau/2}\sqrt{(\mathbb{E}(g))^\tau+\epsilon^\tau} \tag{53}$$

holds true with confidence at least

$$1-\mathscr{N}\left(\mathcal{H},\frac{\epsilon}{|\phi_+'(-1)|}\right)\exp\left\{-\frac{m\epsilon^{2-\tau}}{2\left(C_1+\frac{2}{3}\phi(-1)\epsilon^{1-\tau}\right)}\right\}. \tag{54}$$

Applying the inequality (51) to the upper bound in (53) with $f=\widehat{f}_{\mathbf{z}}\in\mathcal{H}$, we find

$$\begin{aligned}&4\epsilon^{1-\tau/2}\sqrt{\left(\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)\right)^\tau+\epsilon^\tau}\\&\le\frac{\tau}{2}\left(\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)\right)+(1-\tau/2)4^{1/(1-\tau/2)}\epsilon+4\epsilon\\&\le\frac{1}{2}\left(\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)\right)+12\epsilon.\end{aligned} \tag{55}$$

Step 3: The estimation of $\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)$ using (50).

Take $\epsilon=\epsilon^*$. Then the confidence given by (54) is at least $1-\frac{\delta}{2}$. Thus by combining (50), (52), (53) and (55), we know that with confidence at least $1-\delta$,

$$\begin{aligned}\mathcal{E}(\pi(\widehat{f}_{\mathbf{z}}))-\mathcal{E}(f_\rho^\phi)\le4\mathcal{D}(\mathcal{H})+\frac{8C_0'\log\frac{2}{\delta}}{3\mathcal{M}}\\+2\left(\frac{8C_1\log\frac{2}{\delta}}{\mathcal{M}}\right)^{1/(2-\tau)}+24\epsilon^*.\end{aligned}$$

This completes the proof of the theorem. ■

ACKNOWLEDGMENTS

The first and second authors are supported partially by the Research Grants Council of Hong Kong [Projects # CityU 21207019, # CityU 11306220] and by City University of Hong Kong [Project # CityU 7200608]. The last author is supported partially by the Research Grants Council of Hong Kong [Project # CityU 11308020], Hong Kong Institute for Data Science, National Science Foundation of China [Project # 12061160462], and Laboratory for AI-powered Financial Technologies.

REFERENCES

[1] J. Y. Audibert, A. B. Tsybakov, "Fast learning rates for plug-in classifiers", Annals of Statistics, vol.35, no.2, pp.608-633, Apr.2007.
[2] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function", IEEE Trans. Inform. Theory, vol.39, no.3, pp.930-945,May.1993
[3] P. L. Bartlett, M. I. Jordan and J. D. McAuliffe, "Convexity, classification, and risk bounds," J. Amer. Statist. Assoc., vol.101, no.473, pp.138-156, Jan.2006.
[4] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," SIAM Journal on Mathematics of Data Science vol.1, no.1, pp.8-45, Feb.2019.
[5] C. de Boor, G. J. Fix, "Spline approximation by quasiinterpolants," Journal of Approximation Theory, vol.8, pp 19-45, 1973.
[6] G. Brown and F. Dai, "Approximation of smooth functions on compact two-point homogeneous spaces," J. Funct. Anal., vol.220, no.2, pp.401-423, May.2005.
[7] D. R. Chen, Q. Wu, Y. M. Ying, and D. X. Zhou, "Support vector machine soft margin classifiers: Error analysis," J. Machine Learn. Research,vol.5, pp.1143-1175, Sep.2004.
[8] C. K. Chui, S. B. Lin, B. Zhang, and D. X. Zhou, "Realization of spatial sparseness by deep ReLU nets with massive data," IEEE Transactions on Neural Networks and Learning Systems, to be published, DOI:10.1109/TNNLS.2020.3027613.
[9] C. K. Chui, S. B. Lin, D. X. Zhou, "Deep neural networks for rotation-invariance approximation and learning," Analysis and Applications, Vol.17, No.5, pp. 737-772, Aug.2019.
[10] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification," International Conference on Document Analysis and Recognition, Beijing, pp.1135-1139, 2011.
[11] F. Cucker and D. X. Zhou. Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.
[12] R. Durrer. The Cosmic Microwave Background, Cambridge University Press, 2020.
[13] Z. Y. Fang, H. Feng, S. Huang, and D. X. Zhou, "Theory of deep convolutional neural networks II: Spherical analysis," Neural Networks, vol.131, pp.154-162, Nov.2020.
[14] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[15] Z. Han, S. Q. Yu, S. B. Lin, and D. X. Zhou, "Depth selection for deep ReLU nets in feature extraction and generalization," IEEE Transactions on Pattern Analysis and Machine Intelligence, to be published, DOI: 10.1109/TPAMI.2020.3032422.

[16] M. Imaizumi and K. Fukumizu, "Deep neural networks learn non-smooth functions effectively," in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.

[17] Y. D. Kim, I. Ohn, D. Kim, "Fast convergence rates of deep neural networks for classification," Neural Networks,vol.138, pp.179-197, Jun.2021.

[18] J. Klusowski and A. Barron, "Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls," IEEE Transactions on Information Theory **64** pp.7649-7656, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS* (2012): 1097-1105.

[20] J. Lastovicka. "Monitoring and forecasting of ionospheric space weatherłeffects of geomagnetic storms", Journal of Atmospheric and Solar-Terrestrial Physics, vol.64, pp.697-705,2002.

[21] Y. Lin, "Support vector machines and the Bayes rule in classification," Data Mining and Knowledge Discovery, vol.6, pp.259-275, Jul.2002.

[22] Z. Lu, C. Xu, B. Du, T. Ishida, L. Zhang and M. Sugiyama, "LocalDrop: A Hybrid Regularization for Deep Neural Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi:10.1109/TPAMI.2021.3061463.

[23] E. Mammen, and A. Tsybakov, "Smooth discrimination analysis," Annals of Statistics, vol.27, pp.1808-1829, 1999.

[24] T. Mao, Z. J. Shi, and D. X Zhou, "Theory of deep convolutional neural networks III: Approximating radial functions," Neural Networks, to be published.

[25] H. N. Mhaskar, "Approximation properties of a multilayered feedforward artificial neural network," Adv. Comput. Math., vol.1, pp,61-80,1993.

[26] K. Oono and T. Suzuki, "Approximation and non-parametric estimation of ResNet-type convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning (PMLR) 97:4922-4931, 2019.

[27] P. Petersen and V. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep ReLU neural networks," Neural Networks vol.108, pp.296-330, Dec.2018.

[28] J. Schmidt-Hieber, Nonparametric regression using deep neural networks with ReLU activation function, arXiv preprint arXiv: 1708.06633, 2017.

[29] S. Smale and D. X. Zhou, "Learning theory estimates via integral operators and their approximations," Constr. Approx., vol.26, pp.153–172, Apr.2007.

[30] T. Suzuki, Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality, in Proceedings of the International Conference on Learning Representations (ICLR), 2019.

[31] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," Annals of Statistics vol.32, no.1, pp.135-166, Feb.2004.

[32] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37, no.3, pp.328-339, Mar.1989.

[33] Q. Wu, Y. Ying, and D. X. Zhou, "Multi-kernel regularized classifiers," J. Complexity, vol.23, no.1, pp.108-134, Feb.2007.

[34] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," Neural Networks, vol.94, pp.103-114, Oct.2017.

[35] Y. Yang, "Minimax nonparametric classification. I. Rates of convergence." IEEE Transactions on Information Theory 45.7, PP.2271-2284, 1999

[36] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," Annals of Statistics, pp.56-85, 2004.

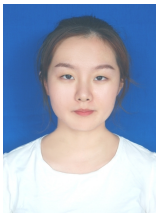
[37] D. X. Zhou, "Universality of deep convolutional neural networks," Appl. Comput. Harmonic Anal., vol.48, no.2, pp.787-794, Mar.2020.

[38] D. X. Zhou, "Theory of deep convolutional neural networks: Downsampling," Neural Networks, vol.124, pp.319-327, Apr.2020.

[39] D. X. Zhou, "Deep distributed convolutional neural networks: universality," Anal. Appl. vol.16, no.6, pp.895-919, 2018, .

[40] T. Y. Zhou and D. X. Zhou, "Theory of deep CNNs induced by 2D convolutions," preprint, 2020.

Han Feng received the B.S.degree in mathematics from Beijing Normal University, China, in 2011, and the M. Sc. and Ph.D. degrees in mathematics from University of Alberta, Canada, in 2013 and 2016, respectively. She is currently an assistant professor in the Department of Mathematics, City University of Hong Kong. Her research interets include spherical approximation theory, learning theory, and deep neural networks.



Shuo Huang received the B.Sc.degree in mathematics from Hohai University, Nanjing, China, in 2019. She is currently a Ph.D student in the Department of Mathematics, City University of Hong Kong. Her research interets include spherical approximation theory, learning theory, and deep neural networks.

Ding-Xuan Zhou received B.Sc. and Ph.D. in mathematics from Zhejiang University, Hangzhou, China, in 1988 and 1991, respectively. He is currently a Chair Professor and associate Dean of the School of Data Science, and Director of Liu Bie Ju Centre for Mathematical Sciences, City University of Hong Kong. His research interests include mathematical theory of deep learning, statistical learning theory, data science, wavelet analysis, and approximation theory.