Approximation with CNNs in Sobolev Space: 
with Applications to Classification

Guohao Shen∗
Department of Applied Mathematics, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong SAR, China
guohao.shen@polyu.edu.hk

Yuling Jiao∗
School of Mathematics and Statistics
and Hubei Key Laboratory of Computational Science
Wuhan University, Wuhan 430072, China
yulingjiaomath@whu.edu.cn

Yuanyuan Lin†
Department of Statistics, The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong SAR, China
ylin@sta.cuhk.edu.hk

Jian Huang†
Department of Applied Mathematics, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong SAR, China
j.huang@polyu.edu.hk

Abstract

We derive a novel approximation error bound with an explicit prefactor for Sobolev-regular functions using deep convolutional neural networks (CNNs). The bound is non-asymptotic in terms of the network depth and filter lengths, in a rather flexible way. For Sobolev-regular functions which can be embedded into the Hölder space, the prefactor of our error bound depends on the ambient dimension polynomially instead of exponentially as in most existing results, which is of independent interest. We also establish a new approximation result when the target function is supported on an approximate lower-dimensional manifold. We apply our results to establish non-asymptotic excess risk bounds for classification using CNNs with convex surrogate losses, including the cross-entropy loss, the hinge loss, the logistic loss, the exponential loss and the least squares loss. We show that the classification methods with CNNs can circumvent the curse of dimensionality if input data is supported on a neighborhood of a low-dimensional manifold.

1 Introduction

Classification methods with hypothesis spaces specified through deep neural networks have achieved remarkable successes in a variety of machine learning tasks [36]. In particular, convolutional neural

∗Guohao Shen and Yuling Jiao contributed equally to this work
†Corresponding authors

networks (CNNs) have demonstrated outstanding performance in many applications, including computer vision \[53\], natural language processing \[57\], and sequence analysis in bioinformatics \[11, 65\]. However, to the best of our knowledge, there have only been limited studies on the approximation properties with CNNs and non-asymptotic excess risk bounds for various classification methods using CNNs.

### 1.1 Related work

There has been much effort devoted to understanding the approximation properties of deep neural networks in recent years. Many interesting results have been obtained concerning the approximation power of deep neural networks for multivariate functions; some examples of important recent works include \[59\], \[60\], \[40\], \[49\], \[53\], \[54\], \[44\], \[13\], \[28\]. These works focused on the approximation power for ReLU activated feedforward neural networks (FNNs) on different kinds of smooth functions.

The approximation power of CNNs has been studied in limited works. \[2\] studied the approximation power of the composed CNNs, where the network consists of fully-connected CNNs followed by a fully-connected layer. They proved that under suitable conditions, convolution neural networks can inherit the universal approximation property of its last fully connected layers. \[48\] showed that for translation equivariant functions, all upper and lower approximation bounds are equivalent between FNNs and CNNs with the same order of network length and size. \[47\] showed that ResNet-type CNNs, a special CNN architecture with skip-layer connections (or identity connections between inconsecutive layers), can replicate the learning ability of FNNs having block-sparse structures. The ResNet-type CNN in \[47\], as claimed, can be dense, and its width, channel size, and filter size are constant with respect to sample size rather than the diverging one in \[48\].

In addition, a series of papers studied the approximation power of deep CNNs with Toeplitz type convolutional matrices \[66\], \[64\], \[63\]. The universality of such CNNs was established in \[64\] for target functions restricted to the Sobolev space \(W^{\beta,2}(\mathbb{R}^d)\) with \(\beta \geq d / 2 + 2\). The approximation error is of the order \(O(L^{-1/2-1/d})\) (up to a logarithmic factor), where \(L\) is the number of layers of such CNNs. But the CNN defined in \[64\] has width increasing linearly with respect to depth. To overcome the difficulty in theoretical analysis of such CNNs, \[63\] introduced a downsampling operator to reduce the widths and the approximation power of downsampled CNNs for a special class of functions (\(\beta\)-Hölder continuous ridge function) was studied. The approximate rate is shown to be \(O(N^{-\beta})\) by downsampled CNNs with a uniform filter length \(s = O(4N + 6)\) and width \(W = O(N)\) for positive integer \(N\). Further, it was shown that a downsampled CNN can compute the same function as a FNN does, with the total number of free parameters of downsampled CNN being at most 8 times of that of FNN. For the CNNs considered by \[66\], \[16\], \[63\] and \[64\], each convolutional layer has only one filter with size \(s\) and the weight matrix is of a special Toeplitz type with shape \((d_{in} + s) \times d_{in}\) and no fully-connected layer is allowed in hidden layers. Approximation properties of convolutional architectures for target functions defined on infinite time domains tailored to temporal sequence data were studied in \[27\].

To mitigate the curse of dimensionality and show the advantages of CNNs in classification problems, several works studied the approximation as well as the statistical learning theories under some low-dimensional assumptions on the target function or the data distribution. In \[32\], \[33\], the posterior probability function is assumed to be Hölder smooth satisfying hierarchical max-pooling model. In \[34\], the target function is assumed to be spatially rotation invariant. In \[18\] and \[48\], input data is assumed to be supported on spheres and manifolds, respectively. Under different structural assumptions on the target function, similar results have also been proved for FNNs \[7\], \[51\] \[52\]. If data is assumed to be supported on low-dimensional manifolds, the curse of dimensionality can also be mitigated for FNNs \[50\], \[44\], \[12\], \[28\].

### 1.2 Our contributions

We study the approximation power of CNNs for functions in Sobolev spaces, including the special case of Hölder class and apply our new approximation results to establish non-asymptotic error bounds for several important classification methods using CNNs. Our main contributions are as follows.

(i) We derive a novel error bound for approximating Sobolev-regular functions using deep CNNs, with the error bound explicitly expressed in terms the network parameters and model parameters.
The error bound depends on the network depth and filter lengths in a rather flexible way. For functions which can be embedded into Hölder space, the prefactor of our error bound depends on the ambient dimension polynomially instead of exponentially as in existing results. We also establish a novel approximation result when the target function is supported on an approximate lower-dimensional manifold, which shows that CNNs are capable of mitigating the curse of dimensionality under such a distributional condition.

(ii) As an application of our CNN approximation results, we establish non-asymptotic bounds for the stochastic and approximation errors of classification methods with a general class of convex surrogate losses in Sobolev space using CNNs. Our main results are also applicable to other problems that use CNNs approximations.

(iii) We apply our general results to establish the non-asymptotic excess risk bounds for classification using CNNs with convex surrogate losses in Sobolev space, including the cross-entropy loss, the hinge loss (SVM), the logistic loss, the exponential loss, and the least squares loss. We show that classification methods with CNNs can circumvent the curse of dimensionality if the input data is supported on a neighborhood of a low-dimensional manifold embedded in the ambient space.

<table>
<thead>
<tr>
<th>Network</th>
<th>Target function</th>
<th>Flexible filter length</th>
<th>Explicit prefactor</th>
<th>Low-dimensional Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>[48] CNN FNN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[47] ConvResNet FNN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[64] CNN Sobolev</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[33] CNN Hölder</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[38] ConvResNet Besov</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

This paper CNN Sobolev and Hölder ✓ ✓ ✓

Table 1: A comparison of some recent convolutional neural network approximation results.

In Table 1 we summarize and compare some recent and most related work on CNN approximation results for various function spaces.

The statistical convergence properties of the excess risk of the empirical risk minimizer (ERM) for the misclassification 0-1 loss have been studied extensively under various conditions. Different types of hypothesis spaces have been used in these approaches, including the linear space, the reproducing kernel Hilbert space and the class of tree-based models, see, for example, [9, 55, 42, 6, 29, 8] and [62]. However, there are limited studies worked on deep binary classifications by FNNs [46, 25, 30], ResNet [26, 45] and CNNs [32, 33, 38].

2 Approximation power of CNNs

2.1 Convolutional neural networks

Convolutional networks are a special type of structured sparse feedforward neural network (FNN) that use convolution in place of general matrix multiplication in at least one of their layers [21]. There are different formulations of CNNs in the literature [2, 66, 47, 37, 63, 64, 33, 32, 38]. In this paper, we consider ReLU activated downsampled CNNs with bias vectors in the convolutional layers defined in [63]. We do not require the norm of network parameters (weight and bias) to be uniformly bounded.

We consider a general CNN function \( f_{\text{CNN}} : \mathcal{X} \to \mathbb{R} \), where \( \mathcal{X} \subset \mathbb{R}^d \). For simplicity, we take \( \mathcal{X} = (0, 1)^d \). Suppose \( f_{\text{CNN}} \) has \( L \) number of hidden layers, then it can be expressed as

\[
f_{\text{CNN}}(x) = A_{L+1} \circ A_L \circ \cdots \circ A_2 \circ A_1(x), x \in \mathcal{X},
\]

where \( \circ \) denotes the functional composition. The \( A_i \)'s are either convolutional operators or down-sampling operators. For convolutional layers, \( A_i(x) = \sigma(W_i^c x + b_i^c) \), where \( W_i^c \in \mathbb{R}^{d_i \times d_{i-1}} \) is the structured sparse Toeplitz type weight matrix induced by the convolutional filter \( \{w_i^{(j)}\}_{j=0}^{s_i^{(j)}} \) with filter length \( s_i^{(j)} \in \mathbb{N}^+ \), and \( b_i^c \in \mathbb{R}^{d_i} \) is a bias vector, and \( \sigma \) is the rectified linear unit (ReLU) activation function applying to each component of the input vector.
For downsampling layers, for any $x \in \mathbb{R}^{d_{i-1}}$, $A_i(x) = D_i(x) = (x_{jm})_{j=1}^{d_i}$, where $D_i : \mathbb{R}^{d_{i-1}} \rightarrow \mathbb{R}^{d_i \times d_{i-1}}$ is the downsampling operator with scaling parameter $m_i \leq d_{i-1}$ in the $i$-th layer. We introduce the convolutional operation and downsampling operation [63] in details in Appendix ??.

We focus on the function class consisting of deep CNNs, denoted by $\mathcal{F}$. We use the Sobolev and Hölder classes for the target functions in our approximation results since $L^p$-smooth functions by networks with squared ReLU activation can be the minimum and maximum filter length over convolutional layers respectively.

**2.2 Approximation in Sobolev space**

We use the Sobolev and Hölder classes for the target functions in our approximation results since they are well-established formulations for describing smooth functions in the literature. Indeed, earlier works on approximation theory of neural networks have been developed for Sobolev functions [59, 64, 24] and for Hölder functions [63, 51]. Let $\mathbb{N}_0$ be the set of non-negative integers and $\beta \in \mathbb{N}_0$. The Sobolev class of functions $W^{\beta,p}(X)$ with $X = (0, 1)^d$ is defined as

$$W^{\beta,p}(X) = \left\{ f \in L^p(X) : D^\alpha f \in L^p(X) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } \|\alpha\|_1 \leq \beta \right\},$$

where $1 \leq p \leq \infty$, $D^\alpha = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$ with $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}_0^d$, and $\|\cdot\|_1 = \sum_{i=1}^d |\alpha_i|$. For $f \in W^{\beta,p}(X)$ and $0 \leq m \leq \beta$, we define the norm

$$\|f\|_{W^{m,p}(X)} := \left( \sum_{0 \leq |\alpha|_1 \leq m} \|D^\alpha f\|_{L^p(X)}^p \right)^{1/p} \quad \text{for } 1 \leq p < \infty,$$

and define $\|f\|_{W^{m,\infty}(X)} := \max_{0 \leq |\alpha|_1 \leq m} \|D^\alpha f\|_{L^\infty(X)}$.

In the remainder of the paper, for any positive integers $M, N \in \mathbb{N}^+$, let $\mathcal{F}_{\text{CNN}}$ be the class of CNNs defined in [2] with depth $L$, size $S$ and filter length specified as follows:

$$L \leq 42(\lfloor \beta \rfloor + 1)^2 M \log_2(8M) \left\lfloor \frac{W-1}{s_{\text{min}}-1} \right\rfloor,$$

$$2 \leq s_{\text{min}} \leq s_{\text{max}} \leq W,$$

$$S \leq 8WL,$$

(7) denotes the maximum incremental width (number of neurons) for consecutive layers in the network and $\lfloor x \rfloor$ denotes the smallest integer no less than $x$.

**Theorem 2.1.** Assume that $f \in W^{\beta,p}(X)$ with $1 \leq \beta \in \mathbb{N}_0$, $1 \leq p \leq \infty$ and $\|f\|_{W^{\beta,p}(X)} \leq B_0$ for $B_0 > 0$. For any $M, N \in \mathbb{N}^+$, there exists a function $f_{\text{CNN}} \in \mathcal{F}_{\text{CNN}}$ defined in [2] with depth $L$ and filter lengths and size $S$ specified in (4), (5) and (6), such that for $m = 0, 1$,

$$\|f - f_{\text{CNN}}\|_{W^{m,p}(X)} \leq C_0(d, \beta, p)(NM)^{-2(\beta - m)/d},$$

where $C_0(d, \beta, p) = 37 \cdot 2^{2\beta+2d/p} B_0^3 (\beta + 1)^3 \times \left\{ \pi^{-d/2} \Gamma(d/2 + 1) \right\}^{2/p+1} (1 + 2\sqrt{d})d^d$. Here $\Gamma(\cdot)$ is the gamma function.

Our result is new for approximating Sobolev-regular functions in $W^{\beta,p}(X)$ for a positive integer $\beta \in \mathbb{N}^+$ with respect to the Sobolev norms $\|\cdot\|_{W^{m,p}(X)}$ for $m = 0, 1$ in terms of explicitly defined approximation error prefactor, clearly defined and flexible network parameters, as well as an fast approximation rate comparable with [22]. A toy example is provided in Appendix D to illustrate the approximation power of CNNs and examine how the approximation error varies according to the filter size and depth of the networks.

In [24], an approximation result on $\beta$-smooth functions by networks with squared ReLU activation function is also provided under the norm $\|\cdot\|_{W^{m,p}(X)}$ for positive integer $m < \beta$. In [22], the
approximation rates of ReLU DNN for Sobolev-regular functions with respect to the weaker Sobolev norms \( \| \cdot \|_{W^{m,p}(X)} \) for \( s \in [0,1] \) were analyzed, which can be seen as a generalization of the result in [59]. Approximation results on Sobolev-regular functions \( W^{\beta,p}(X) \) by networks with general activation functions are derived in [28], where the error is with respect to \( \| \cdot \|_{W^{m,p}(X)} \) for some integer \( m < \beta \).

By the general Sobolev inequality, see Theorem 6 in Chapter 5 of [13], \( W^{k,p}(X) \) with \( kp > d \) can be embedded into a Hölder class \( \mathcal{H}^\beta(X, B_0) \) (with \( \beta = k - \lfloor d/p \rfloor - 1 \)) defined as

\[
\mathcal{H}^\beta(X, B_0) = \left\{ f : X \to \mathbb{R}, \max_{\|\alpha\|_1 \leq \lfloor \beta \rfloor} \|D^\alpha f\|_\infty \leq B_0 \text{ and } \max_{\|\alpha\|_1 = s, x \neq y} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x - y\|_2} \leq B_0 \right\},
\]

where \( \beta > 0 \) and \( \lfloor \beta \rfloor \in \mathbb{N}_0 \) denotes the largest integer strictly smaller than \( \beta \). Because of the extra regularity on the Hölder class, a much improved prefactor \( C_0 \) in the error bound of Theorem 2.1 can be obtained.

**Theorem 2.2.** Let \( f \in \mathcal{H}^\beta(X, B_0) \) be defined in (8) and let \( X \in \mathbb{R}^d \) be a random vector whose probability distribution is supported on \( X = (0,1]^d \) and absolutely continuous with respect to the Lebesgue measure. For any \( M, N \in \mathbb{N}^+ \), there exists a function \( f_{\text{CNN}} \in \mathcal{F}_{\text{CNN}} \) defined in (2) with depth \( L \) and filter lengths and size \( S \) specified in (4), (5) and (6), such that

\[
\mathbb{E}|f(X) - f_{\text{CNN}}(X)| \leq C_0(d, \beta)(NM)^{-\beta/d},
\]

where \( C_0(d, \beta) = 18B_0(\beta + 1)^2d^{\beta + (\beta + 1)/2} \). Here \( a \vee b := \max\{a, b\} \) for \( a, b \in \mathbb{R} \).

The convergence rate \((NM)^{-\beta/d}\) in Theorems 2.1 and 2.2 with respect to the network depth and filter lengths specified by \( M \) and \( N \), is in line with the nearly optimal rate of ReLU FNNs on smooth functions under Sobolev norm in [24] and on Hölder smooth functions under \( L^\infty \) norm in [28]. The prefactor \( 18B_0(\beta + 1)^2d^{\beta + (\beta + 1)/2} \) of approximation error depends on the dimension \( d \) polynomially, different from the exponential dependence in many existing neural network approximation results mentioned in Section 1.1.

### 2.3 Approximation with a lower-dimensional support

In many modern machine learning problems, the ambient dimension \( d \) of the input data could be very large, which results in an extremely slow convergence rate. This fact is known as the *curse of dimensionality*. Fortunately, many types of data have a low-dimensional latent structure, that is, although the ambient dimension \( d \) is large, the distribution of the data is approximately supported on a low-dimensional subset of \( \mathbb{R}^d \), in which case the approximation error bound can be substantially improved. We establish an approximation result for CNNs in Sobolev spaces under a low-dimensional support assumption.

**Assumption 2.3.** The distribution of \( X \) is supported on \( M_\rho \), a \( \rho \)-neighborhood of \( M \subset X \), where \( M \) is a compact \( d_M \)-dimensional Riemannian submanifold and \( M_\rho = \{ x \in X : \inf\{\|x - y\|_2 : y \in M\} \leq \rho \} \) for \( \rho \in (0,1) \).

In real-world applications, data are hardly observed to locate on an *exact manifold*, instead they could be more realistically viewed as consisting of a latent part supported on a low-dimensional manifold \( M \) plus noises. Therefore, Assumption 2.3 is more realistic than the exact manifold assumption [50] [12] [38].

Define

\[
\Delta = O(d_M \varepsilon^{-2} \log(d/\varepsilon)), \quad \varepsilon \in (0,1), \quad (9)
\]

\[
\rho = C_2(NM)^{-\beta/d}(\beta + 1)^2d^{\beta/2} \varepsilon \times [(d/d_\varepsilon)^{1/2} + 1 - \varepsilon]^{-1} (1 - \varepsilon)^{1-\beta}. \quad (10)
\]

In our error bound results below, it is the \( \Delta \) that will affect the convergence rate. Since it is often the case that \( d_M \ll d \) and therefore \( d_\varepsilon \ll d \), the manifold assumption will lead to a better convergence rate than those in Theorems 2.1 and 2.2. In addition, we require \( \rho \leq \rho_\varepsilon \).

**Theorem 2.4.** Suppose that Assumption 2.3 holds, \( f \in \mathcal{H}^\beta(X, B_0) \) and the distribution of \( X \) is absolutely continuous with respect to the Lebesgue measure. Let \( \Delta_\varepsilon \) and \( \rho_\varepsilon \) be defined in (9) and (10).
and (10), respectively. Assume that \( \rho \leq \rho_2 \). Then, for any \( M, N \in \mathbb{N}^+ \), there exists a CNN function \( f_{\text{CNN}} \in \mathcal{F}_{\text{CNN}} \) with depth \( L \) and filter lengths and size \( S \) specified in (4), (5) and (6) with \( W = 38^2([\beta]+1)^4d^2\beta \beta + 2N^2[\log_2(8N)]^2 \) such that

\[
\mathbb{E}|f(X) - f_{\text{CNN}}(X)| \leq C(d, \beta)(NM)^{-2\beta/d_\epsilon},
\]

where \( C(d, \beta) = (18 + C_2)B_0(1 - \epsilon)^{-\beta}(\beta + 1)^2d_\epsilon^1/2d_\epsilon^{3/2}/2 \).

Since the intrinsic dimension \( d_\epsilon \) can be much smaller than the ambient dimension \( d \), the rate \((NM)^{-2\beta/d_\epsilon}\) in Theorem 2.4 is greatly improved compared with the rate \((NM)^{-2\beta/d}\) in Theorem 2.2 and the curse of dimensionality is mitigated. The result here is of independent interest and can be useful in other problems that involve the use of CNNs.

3 Excess risk in classification

In this section, we present the application of the approximation results Theorems 2.1 and 2.4 to the error analysis of classification with CNNs.

Consider a binary classification problem with a predictor \( X \in \mathcal{X} \subset \mathbb{R}^d \) and its binary label \( Y \in \{1, -1\} \). We are interested in learning a classifier \( h : \mathcal{X} \to \{1, -1\} \) from a class of functions, or a hypothesis space, denoted by \( \mathcal{H} \). Let the joint distribution of \((X, Y)\) be \( \mathbb{P} \). The goal is to find a classifier that minimizes the misclassification error or the 0-1 risk: \( R_*(h) = \mathbb{P}\{h(X) \neq Y\}, h \in \mathcal{H} \).

Denote the misclassification risk minimizer at the population level by \( h_0 = \arg\min_{h \text{ measurable}} R_*(h) \).

For any \( h \in \mathcal{H} \), the excess risk of \( h \) is \( R_*(h) - R_*(h_0) \), the difference between the misclassification errors of \( h \) and \( h_0 \).

Since the probability measure \( \mathbb{P} \) is unknown in practice, the classifier \( h \) will be learned based on a random sample \( S = \{(X_i, Y_i)\}_{i=1}^n \) from \( \mathbb{P} \), where \( n \) is the sample size. The empirical risk minimizer (ERM) is defined by

\[
\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n 1(h(X_i) \neq Y_i).
\]

However, the empirical risk function based on the 0-1 loss is non-continuous and non-convex, thus this minimization problem is typically computationally intractable.

Rather than minimizing the non-smooth 0-1 loss, many popular methods adopt a proper convex loss function to train classifiers with computational efficiency that can be done in polynomial time. Moreover, proper surrogate convex loss functions have been shown to be consistent with the 0-1 loss function for binary classification problems [61, 5].

3.1 Convex surrogate loss functions

Let \( \phi \) be a given convex univariate loss function \( \phi : \mathbb{R} \to [0, \infty) \). We consider the risk function with respect to the loss \( \phi \)

\[
R(f) \equiv R_\phi(f) = \mathbb{E}\{\phi(Yf(X))\},
\]

where for simplicity of notation and without causing confusion we omit the superscript \( \phi \) in \( R \), similarly for \( \hat{f}_n \) and \( f_0 \) defined below. For a given random sample \( S = \{(X_i, Y_i)\}_{i=1}^n \), we denote the ERM over \( \mathcal{F}_{\text{CNN}} \) with a given loss \( \phi \) by

\[
\hat{f}_n \in \arg\min_{f \in \mathcal{F}_{\text{CNN}}} \frac{1}{n} \sum_{i=1}^n \phi(Yf(X_i)).
\]

Based on the ERM \( \hat{f}_n \), a classifier \( \hat{h}_n(x) := \text{sign}(\hat{f}_n(x)) \) for \( x \in \mathcal{X} \) can be defined. As shown in [61, 5], for a properly chosen \( \phi \), \( \hat{f}_n \) can help reduce the excess risk \( R_*(\hat{h}_n) - R_*(h_0) \). Specifically, define the measurable minimizer of \( R \) in (12) as

\[
f_0 = \arg\min_{f \text{ measurable}} \mathbb{E}\{\phi(Yf(X))\},
\]
We bound the stochastic error in terms of the pseudo-dimension of the downsampled CNNs defined in Theorem 4.2. Particularly, we further bound its pseudo-dimension and express the bound in terms of quantities related to CNNs.

### 4 Non-asymptotic error bounds

For the ERM $\hat{f}_n$ defined in [13], we first state a basic inequality for bounding the excess risk of $\hat{f}_n$.

**Lemma 4.1.** For any loss $\phi$ and any random sample $S = \{(X_i, Y_i)\}_{i=1}^n$, the excess $\phi$-risk of the ERM $\hat{f}_n$ satisfies $R(\hat{f}_n) - R(f_0) \leq \text{StoErr} + \text{AppErr}$, where

$$\text{StoErr} = 2\sup_{f \in \mathcal{F}} |R(f) - R_n(f)|, \quad \text{AppErr} = \inf_{f \in \mathcal{F}} R(f) - R(f_0).$$

(16)

Therefore, the excess risk of $\hat{f}_n$ is bounded by the sum of two terms: the stochastic error and the approximation error. For a given loss function $\phi$, the upper bound no longer depends on the ERM $\hat{f}_n$, but the function class $\mathcal{F}$ and the random sample $S$. The stochastic error in (16) depends on the complexity of $\mathcal{F}$ and the approximation error in (16) depends on the approximation power of the class $\mathcal{F}$ for $f_0$.

### 4.1 Stochastic error

We bound the stochastic error in terms of the pseudo-dimension of the downsampled CNNs defined in [2]; particularly, we further bound its pseudo-dimension and express the bound in terms of quantities related to CNNs.

**Theorem 4.2.** [Stochastic error bound] Suppose that $\phi$ is convex and non-increasing. For any $M, N \in \mathbb{N}^+$, let $\mathcal{F}_{\text{CNN}}$ be the class of CNNs defined in [2] with $B \geq 0$, depth $L$ and size $S$ and let $\phi_B := \sup_{|a| \leq B} \phi(a)$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the stochastic error in (16) satisfies

$$\text{StoErr} \leq \frac{2\phi_B}{\sqrt{n}} \left(C_0 \sqrt{SL \log(S) \log(n)} + \sqrt{\log(1/\delta)}\right),$$

where $C_0 > 0$ is a universal constant.
The quantity $\phi_B$ for common loss functions are presented in Table 2. It worth noting that the error bound here does not require the norm of the CNN parameters (of weight and bias) to be uniformly bounded. In comparison, those stochastic error bounds based on the covering number generally assume an uniformly bounded norm on CNN parameters, which may hinder the approximation power of network since most approximation results need the norm of network parameters to tend to infinity as the approximation error tends to zero.

4.2 Approximation error

We derive an upper bound of the approximation error by relating

### Assumption 4.3 (a) and (b) are regular conditions ensuring that the target function $f$ is continu-
ous and the loss function $\phi$ is Lipschitz on bounded intervals. The absolute continuity assumption of the probability measure of $X$ with respect to the Lebesgue measure is reasonable for the approximate low-dimensional manifold case but inapplicable with the exact low-dimensional manifold condition. This assumption is for deriving better $L_p$ approximation error bound, without which the error bound in term of the $L_\infty$ norm can still be established, but at the price of a wider neural network and a larger prefactor. It can be verified that commonly used loss functions, such as the cross entropy, the logistic and the exponential loss functions, satisfy Assumption 4.3(b).

To deal with the approximation of unbounded target function, we truncate the target $f_0$ by a constant $T$, where $T$ may depend on $n$. Let $f_{0,T}$ be the truncated version of $f_0$ defined as

$$f_{0,T}(x) = \begin{cases} 
  f_0(x), & \text{if } |f_0(x)| \leq T, \\
  T \text{sign}(f_0(x)), & \text{if } |f_0(x)| > T.
\end{cases}$$

Denote the error of the loss function $\phi$ due to truncation by

$$\Delta_\phi(T) := \inf_{|a| \leq T} \phi(a) - \inf_{a \in \text{Ran}(f_0)} \phi(a),$$

where $\text{Ran}(f_0)$ is the range of the target function $f_0$. This error decreases as $T$ increases. Then, the approximation error can be decomposed into two terms that are easier to deal with.

**Lemma 4.4.** [ϕ-approximation error] Suppose that Assumption 4.3 holds and $T \leq B$. Then, the ϕ-approximation error $\inf_{f \in F_{\text{CNN}}} R(f) - R(f_0)$ with respect to the loss function $\phi$ satisfies

$$\inf_{f \in F_{\text{CNN}}} R(f) - R(f_0) \leq B_\phi \inf_{f \in F_{\text{CNN}}} E|f(X) - f_{0,T}(X)| + \Delta_\phi(T).$$

where $B_\phi$ is defined as the Lipschitz constant of $\phi$ on the interval $[-B, B]$.

We list $B_\phi$ and $\Delta_\phi(T)$ for hinge and cross entropy loss functions in Table 2. A detailed table including other loss functions are given in the appendix.

**Theorem 4.5.** [Approximation error bound] Suppose that Assumption 4.3 holds and $f_0 \in H^d([0,1]^d, B_0)$. For any $M, N \in \mathbb{N}^+$, let $F_{\text{CNN}}$ be the class of CNNs defined in (2) with $T \leq B$, depth $L$ and filter lengths specified in Table 2 and (3). Then, the approximation error defined in (10) satisfies

$$\text{AppErr} \leq C(d, \beta)(NM)^{-2\beta/5} + \Delta_\phi(T),$$

where $C(d, \beta) = 18B_\phi B_0(\beta + 1)^2a^{3\beta+1}(\beta+1)!/\sqrt{2}$. There is a trade-off since $T \leq B$ and $\Delta_\phi(T)$ decreases in $T$ but $C(d, \beta)$ increases in $B_\phi$ (or $B$).
4.3 Non-asymptotic excess risk bound

Combining Theorems 4.2 and 4.5, we obtain the excess error bound.

**Theorem 4.6.** [Non-asymptotic excess ϕ-risk bound] Suppose that Assumption 4.3 holds and \( f_0 \in \mathcal{H}^F([0, 1]^d, B_0) \). For any \( M, N \in \mathbb{N}^+ \), let \( \mathcal{F}_{\text{CNN}} \) be the class of CNNs defined in (2) with \( T \leq B \), depth \( L \) and filter lengths specified in (4) and (5). Then, for any \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \), the ERM \( \hat{f}_n \), defined in (13), satisfies

\[
R(\hat{f}_n) - R(f_0) \leq \text{StoErr} + \text{AppErr},
\]

with

\[
\text{StoErr} = \frac{2\phi_n}{\sqrt{n}} \left( C_0 \sqrt{SL \log(S) \log(n)} + \sqrt{\log(1/\delta)} \right),
\]

\[
\text{AppErr} = C(d, \beta)(NM)^{1/2} + \Delta_\phi(T),
\]

where \( C(d, \beta) = 18B_0B_0(\beta + 1)^2d^3 + \delta^{3/2}\beta^{1/2} \) and \( C_0 > 0 \) is a universal constant.

The upper bound of the excess ϕ-risk \( R(\hat{f}_n) - R(f_0) \) in Theorem 4.6 is a sum of two error terms, the stochastic error and the approximation error. To achieve the optimal error rate, we need to balance the trade-off between them. On one hand, the bound for the stochastic error \( \text{StoErr} \) increases with the size and the depth of \( \mathcal{F}_{\text{CNN}} \). On the other hand, the bound for the approximation error \( \text{AppErr} \) decreases with the depth and the upper bound of filter lengths of \( \mathcal{F}_{\text{CNN}} \).

4.4 Circumventing the curse of dimensionality

We state a theorem that provides a non-asymptotic excess risk bound under the approximate low-dimensional manifold assumption.

**Theorem 4.7** (Circumventing the curse of dimensionality). Suppose that Assumptions 2, 3, and 4.3 hold, and \( f_0 \in \mathcal{H}^F([0, 1]^d, B_0) \). For any \( M, N \in \mathbb{N}^+ \), let \( \mathcal{F}_{\text{CNN}} \) be the class of CNNs defined in (2) with \( T \leq B \), depth \( L \) and filter lengths and size \( S \) specified in (4), (5), and (6) with \( W = 38^2(\beta + 1)^4d^2N^2(\log_2(8N))^2 \). Suppose \( \rho \) in Assumption 2 holds satisfies \( \rho \leq \rho_\varepsilon \), where \( \rho_\varepsilon \) is defined in (10). Then, for any \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \), the ERM \( \hat{f}_n \), defined in (13), satisfies

\[
R(\hat{f}_n) - R(f_0) \leq \text{StoErr}_\varepsilon + \text{AppErr}_\varepsilon,
\]

with

\[
\text{StoErr}_\varepsilon = \frac{2\phi_n}{\sqrt{n}} \left( C_1 \sqrt{SL \log(S) \log(n)} + \sqrt{\log(1/\delta)} \right),
\]

\[
\text{AppErr}_\varepsilon = \left( 1 + C_2 \right) \frac{B_0B_0}{(1 - \varepsilon)^3} (\beta + 1)^2d^3 + \delta^{3/2}\beta^{1/2}\right) \times (NM)^{1/2} + \Delta_\phi\varepsilon(T),
\]

where \( C_1, C_2 > 0 \) are universal constants.

For a high-dimensional input \( X \) with a large \( d, d_\varepsilon \) satisfies \( d_M \leq d_\varepsilon \leq d \) for \( \varepsilon \in (0, 1) \). For the stochastic error \( \text{StoErr}_\varepsilon \), comparing with the bound (18) in Theorem 4.6, we see that \( \text{StoErr}_\varepsilon = \text{StoErr} \), that is, the stochastic error does not change under the approximate low-dimensional manifold assumption. For the approximation error \( \text{AppErr}_\varepsilon \), we see that the convergence rate \( (NM)^{1/2} \) only depends on \( d_\varepsilon \), which leads to a much faster convergence rate.

5 Examples

In this section, we illustrate the applications of Theorems 4.6 and 4.7 to obtaining non-asymptotic error bounds for the excess risk in classification. We apply the general excess risk bounds established in these theorems to several important classification methods with CNNs when a specific form of \( \phi \) is given. We present the results for the cross entropy loss below. The non-asymptotic error bounds for the hinge, the logistic, the exponential and the least squares losses are given in Appendix ??.

For the cross entropy loss \( \phi(a) = -\log(0.5 + a) \), the Lipschitz constant \( B_\phi \) on \([-0.5, 0.5] \) is not bounded. As the minimizer of \( \phi \)-risk \( \eta = 0.5 \) is bounded, we can choose \( T = B = 0.5 - \eta^{2/2d+4\beta} \). Then \( \Delta_\phi(T) \leq n^{-1/2(2d+4\beta)} \) and \( B_\phi = n^{3/2(2d+4\beta)} \).
Denote $\xi_n = n^{d/(2d+4\beta)}$. Take $N = 1$ in \textcolor{red}{[7]} and $M = |\xi|$ in \textcolor{red}{[4]} and let $\mathcal{F}_{\text{CNN}}$ be the class of CNNs in \textcolor{red}{[2]} with depth $L \leq 378 \cdot 38^2(|\beta| + 1)9d^{2[\beta]+2}[\xi_n] |\log_2(8|\xi_n|)|/(s_{\min} - 1)$, filter lengths $2 \leq s_{\min} \leq s_{\max} \leq 9 \times 38^2(|\beta| + 1)^4d^{2[\beta]+2}$ and size $S \leq 8WL \leq 42 \ast 8 \ast 9^2 \ast 38^4(|\beta| + 1)^{10}d^{4[\beta]+4}[\xi_n] |\log_2(8|\xi_n|)|/(s_{\min} - 1)$. Theorem \textcolor{red}{4.6} implies that, with probability at least $1 - \exp(\xi_n^2)$, the excess $\phi$-risk of the ERM $\hat{f}_n$ defined in \textcolor{red}{[13]} satisfies

$$R(\hat{f}_n) - R(f_0) \leq C n^{-\beta/(2d+4\beta)} (\log n)^2,$$

where $C = O((Bo_1)^8 d^{3\beta+3}/(s_{\min} - 1))$ is a constant independent of $n$.

Let $T = B = 0.5 - (\log n)^{-1}$. Using a modified cross entropy loss $\phi(a) = \max\{-\log(0.5 + a), \tau\}$ with $\tau = -\log(1 - (\log n)^{-1})$, the excess $\phi$-risk of the ERM $f_n$ can be improved to $O(n^{-\beta/(d+2\beta)} (\log n)^2)$.

If the approximate low-dimensional manifold Assumption \textcolor{red}{2.3} holds and for any $\varepsilon \in (0, 1)$, the radius of the neighborhood $\rho$ in Assumption \textcolor{red}{2.3} satisfies $\rho \leq \rho_\varepsilon$, where $\rho_\varepsilon$ is defined in \textcolor{red}{[10]}, then Theorem \textcolor{red}{4.7} implies that the rate of convergence can be improved:

$$R(\hat{f}_n) - R(f_0) \leq C n^{-\beta/(2d+4\beta)} (\log n)^2,$$

where $C$ is a constant independent of $n$.

The above discussion shows that deep binary classifications with CNNs are adaptive to the low dimensional structure of the data and the smoothness of the target. Moreover, the prefactor $C$ depends on the ambient dimension $d$ polynomially, which improves the prefactors depending on $d$ exponentially in the existing excess risk bounds \textcolor{red}{[46, 25, 30, 26, 45, 32, 33, 38]}. Moreover, the filter lengths here are more flexible comparing with those requiring certain filter lengths in \textcolor{red}{[32, 33, 48]}.

### 6 Conclusion

In this work, we derive new approximation error bounds with explicit prefactor in terms of network parameters for Sobolev-regular functions and Hölder smooth functions using deep convolutional networks. New approximation result when the target function is supported on an approximate lower-dimensional manifold is established. Different from existing results, the prefactor of our error bound depends on the ambient dimension polynomially instead of exponentially for Hölder smooth functions. The new approximation results are applied to establish non-asymptotic excess risk bounds for a class of classification methods using CNNs.

An important limitation of our work is that we only considered binary classification problems in applying our approximation results using deep convolutional networks. It would be interesting to apply the results in this work to other settings involving CNN approximation, such as the multiclass classification and regression problems. In addition, our work only partially explains the empirical successes of CNNs in practice. For image data, the assumption of approximate lower-dimensional support does not capture all the structural information. For example, spatial invariance is likely to be expected in some problems such as image classification. We believe that if such properties are taken into account, the theoretical bounds can be further improved. We hope to study this in the future.

### Acknowledgements

The authors wish to thank three anonymous reviewers for their comments and suggestions that helped improve the paper significantly.

G. Shen is partially supported by the research grant P0041243 from The Hong Kong Polytechnic University. Y. Jiao is supported in part by the National Science Foundation of China under Grant 11871474, the research fund of KLATASDSMOE, and the Fundamental Research Funds for the Central Universities NO. 2042022K0071. Y. Lin is supported by the Hong Kong Research Grants Council (Grant Nos. 14306219 and 14306620), the National Natural Science Foundation of China (Grant No. 11961028) and Direct Grants for Research, The Chinese University of Hong Kong. J. Huang is partially supported by the research grant P0042888 from The Hong Kong Polytechnic University.
References


