Approximation and Non-parametric Estimation of ResNet-type Convolutional Neural Networks

Kenta Oono [1 2] **Taiji Suzuki** [1 3]

Abstract

Convolutional neural networks (CNNs) have been shown to achieve optimal approximation and estimation error rates (in minimax sense) in several function classes. However, previous analyzed optimal CNNs are unrealistically wide and difficult to obtain via optimization due to sparse constraints in important function classes, including the Hölder class. We show a ResNet-type CNN can attain the minimax optimal error rates in these classes in more plausible situations – it can be dense, and its width, channel size, and filter size are constant with respect to sample size. The key idea is that we can replicate the learning ability of Fully-connected neural networks (FNNs) by tailored CNNs, as long as the FNNs have *block-sparse* structures. Our theory is general in a sense that we can automatically translate any approximation rate achieved by block-sparse FNNs into that by CNNs. As an application, we derive approximation and estimation error rates of the aformentioned type of CNNs for the Barron and Hölder classes with the same strategy.

1. Introduction

Convolutional neural network (CNN) is one of the most popular architectures in deep learning research, with various applications such as computer vision (Krizhevsky et al., 2012), natural language processing (Wu et al., 2016), and sequence analysis in bioinformatics (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015). Despite practical popularity, theoretical justification for the power of CNNs is still scarce from the viewpoint of statistical learning theory.

For fully-connected neural networks (FNNs), there is a lot of existing work, dating back to the 80's, for theoretical explanation regarding their *approximation* ability (Cybenko, 1989; Barron, 1993; Lu et al., 2017; Yarotsky, 2017; Lee et al., 2017; Petersen & Voigtlaender, 2018b) and *generalization* power (Barron, 1994; Arora et al., 2018; Suzuki, 2018). See also surveys of earlier work by Pinkus (2005) and Kainen et al. (2013). Although less common compared to FNNs, recently, statistical learning theories for CNNs have been studied both about approximation ability (Zhou, 2018; Yarotsky, 2018; Petersen & Voigtlaender, 2018a) and generalization power (Zhou & Feng, 2018). Among others, Petersen and Voigtlaender (2018a) showed any function realizable by an FNN is representable with an (equivariant) CNN that has the same order of parameters. This fact means virtually any approximation and estimation error rates achieved by FNNs can be achieved by CNNs, too. In particular, because FNNs are optimal in minimax sense (Tsybakov, 2008; Giné & Nickl, 2015) for several important function classes such as the Hölder class (Yarotsky, 2017; Schmidt-Hieber, 2017), CNNs are also minimax optimal for these classes.

However, the optimal CNN obtained by the result of (Petersen & Voigtlaender, 2018b) can be unrealistically *wide*: for D variate β-Hölder case (see Definition 4), its depth is $O(\log N)$, while its channel size is as large as $O(N^{\frac{D}{2\beta+D}})$ where N is sample size. To the best of our knowledge, no CNNs that achieve the minimax optimal rate in important function classes, including the Hölder class, can keep the number of units per layer constant with respect to N. Thanks to recent techniques such as identity mappings (He et al., 2016; Huang et al., 2018), sophisticated initialization schemes (He et al., 2015; Chen et al., 2018), and normalization methods (Ioffe & Szegedy, 2015; Miyato et al., 2018), architectures that are considerably deep and moderate channel size and width have become feasible. Therefore, we would argue that there are growing demands for theories which can accommodate such constant-size architectures.

The other issue is impractical *sparsity* constraints imposed on neural networks. Existing literature (Schmidt-Hieber, 2017; Suzuki, 2019; Imaizumi & Fukumizu, 2019) proved the minimax optimal property of FNNs for several func-

[1] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan [2] Preferred Networks, Inc. (PFN), Tokyo, Japan [3] Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. Correspondence to: Kenta Oono <kenta_oono@mist.i.u-tokyo.ac.jp>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

tion classes. However, they picked an estimator from a set of functions realizable by FNNs with a given number of non-zero parameters. For example, Schmidt-Hieber (2017) constructed an optimal FNN that has depth $O(\log N)$, width $O(N^{\alpha})$, and $O(N^{\alpha}\log N)$ non-zero parameters when the true function is D variate β-Hölder. Here, N is the sample size and $\alpha = \frac{D}{2\beta+D}$. It means the ratio of non-zero parameters (i.e., the number of non-zero parameters divided by the number of all parameters) is $\tilde{O}(N^{-\alpha})$. To obtain such neural networks, we need to consider impractical combinatorial problems such as L_0 norm optimization. Although we can obtain minimax optimal CNNs using the equivalence of CNNs and FNNs explained before, these CNNs have the same order of sparsity, too.

In this paper, we show that CNNs can achieve minimax optimal approximation and estimation error rates, even they have more plausible architectures. Specifically, we analyze the learning ability of ResNet-type (He et al., 2016) CNNs with ReLU activation functions (Krizhevsky et al., 2012) which can be dense and have constant width, channel size, and filter size against the sample size. There are mainly two reasons that motivate us to study this type of CNNs. First, although ResNet is a de facto architecture in various practical applications, the minimax optimal property for ResNet has not been explored extensively. Second, constant-width CNNs are critical building blocks not only in ResNet but also in various modern CNNs such as Inception (Szegedy et al., 2015), DenseNet (Huang et al., 2017), and U-Net (Ronneberger et al., 2015), to name a few.

Our strategy is to emulate FNNs by constructing tailored ResNet-type CNNs in a similar spirit to Zhou (2018) and Petersen and Voigtlaender (2018a). The unique point of our method is to pay attention to a *block-sparse* structure of an FNN, which roughly means a linear combination of multiple possibly dense FNNs. Block-sparseness decreases the model complexity coming from the combinatorial sparsity patterns and promotes better bounds. Therefore, approximation and learning theories of FNNs often utilized it both implicitly or explicitly (Yarotsky, 2018; Bölcskei et al., 2019). We first prove that if an FNN is block-sparse with M blocks, we can realize the FNN with a ResNet-type CNN with $O(M)$ additional parameters. In particular, if blocks in the FNN are dense, which is often true in typical settings, increase of parameters in number is negligible. Therefore, the order of approximation rate of CNNs is same as that of FNNs, and hence we can also show that the CNNs can achieve the same estimation error rate as the FNNs. We also note that the CNN does not have sparse structures in general in this case. Although our primary interest is the Hölder class, this result is general in the sense that it is not restricted to a specific function class, as long as we can approximate it using block-sparse FNNs.

To demonstrate the broad applicability of our methods, we derive approximation and estimation errors for two types of function classes with the same strategy: the Barron class (of parameter $s = 2$, see Definition 3) and Hölder class. We prove, as corollaries, that our CNNs can achieve the approximation error of order $\tilde{O}(M^{-\frac{D+2}{2D}})$ for the Barron class and $\tilde{O}(M^{-\frac{\beta}{D}})$ for the β-Hölder class and the estimation error of order $\tilde{O}_P(N^{-\frac{D+2}{2(D+1)}})$ for the Barron class and $\tilde{O}_P(N^{-\frac{2\beta}{2\beta+D}})$ for the β-Hölder class, where M is the number of parameters (we used M, which is same as the number of blocks, to indicate the parameter count because it will turn out that CNNs have $\Omega(M)$ blocks for these cases), N is the sample size, and D is the input dimension. These rates are same as the ones for FNNs ever known in existing literature. An important consequence of our theory is that the ResNet-type CNN can achieve the minimax optimal estimation error (up to logarithmic factors) for the Hölder class even if it can be dense, and its width, filter size, and channel size are constant against sample size. This fact is in contrast to existing work, where optimal FNNs or CNNs are inevitably sparse and have width or channel size going to infinity as $N \to \infty$. Further, we prove minimax optimal CNNs can have constant-depth residual blocks for the Hölder case, if we introduce signal scaling mechanisms to CNNs (see Definition 5).

In summary, the contributions of our work are as follows:

- We develop general approximation theories for CNNs via ResNet-type architectures. If we can approximate a function with a block-sparse FNN with M dense blocks, we can approximate the function with a ResNet-type CNN at the same rate, too (Theorem 1). The CNN is dense in general and is not assumed to have unrealistic sparse structures.

- We derive the upper bound of the estimation error of ResNet-type CNNs (Theorem 2). It gives a sufficient condition to obtain the same estimation error rate as that of FNNs (Corollary 1).

- We apply our theory to the Barron and Hölder classes and derive the approximation (Corollary 2 and 4) and estimation (Corollary 3 and 5) error rates, which are identical to those for FNNs, even if the CNNs are dense and have constant width, channel size, and filter size with respect to sample size. This rate is minimax optimal for the Hölder case.

- For the Hölder case, the optimal CNNs can additionally have constant-depth residual blocks if we introduce scaling mechanism to identity mappings (Theorem 3 and 4).

2. Related Work

In Table 1, we highlight differences in CNN architectures between our work and work done Zhou (2018) and Petersen and Voigtlaender (2018a), which established approximation theories of CNNs via FNNs.

First and foremost, Zhou only considered a specific function class — the Barron class — as a target function class, although we can apply their method to any function class realizable by a 2-layered ReLU FNN (i.e., a ReLU FNN with a single hidden layer). Regarding architectures, they considered CNNs with a single channel and whose width is "linearly increasing" (Zhou, 2018) layer by layer. For regression or classification problems, it is rare to use such an architecture. Besides, since they did not bound the norm of parameters in approximating CNNs, we cannot derive the estimation error from their result.

Petersen and Voigtlaender (2018a) fully utilized a group invariance structure of underlying input spaces to construct CNNs. Such a structure makes theoretical analysis easier, especially for investigating the equivariance properties of CNNs because it enables us to incorporate mathematical tools such as group theory, Fourier analysis, and representation theory (Cohen et al., 2018). Although their results are quite general in a sense that we can apply it to any function that can be approximated by FNNs, their assumption on group structures excludes the padding convolution layer, a popular type of convolution operations. Secondly, if we simply combine their result with the approximation result of Yarotsky (2017), the CNN which optimally approximates β-Hölder function by the accuracy ε (with respect to the sup-norm) has $\tilde{O}(\varepsilon^{-\frac{D}{\beta}})$ channels, which grows as $\varepsilon \to 0$ (D is the input dimension). Finally, the ratio of non-zero parameters of optimal CNNs is $\tilde{O}(N^{-\frac{D}{2\beta+D}})$. That means the optimal CNNs gets incredibly sparse as the sample size N increases. One of the reasons for the large channel size and sparse structure is that their construction was not aware of the sparse internal structure of approximating FNNs, which motivates us to consider special structures of FNNs, the block-sparse structure.

As opposed to these two studies, we employ padding- and ResNet-type CNNs which have multiple channels, fixed-sized filters, and constant width. Like Petersen and Voigtlaender (2018a), we can apply our result to any function, as long as FNNs to be approximated are block-sparse, including the Barron and Hölder cases. If we apply our theorem to these classes, we can show that the optimal CNNs can achieve the same approximation and estimation rates as FNNs, while they are dense and the number of channels is independent of the sample size.

Finite-width neural networks have been studied in earlier work (Lu et al., 2017; Perekrestenko et al., 2018; Fan et al., 2018). However, they only derived approximation abilities. For finite-width networks, it is far from trivial to derive optimal estimation error rates from approximation results: if a network approximates a true function more accurately while restricting its capacity per layer, the neural network inevitably gets deeper. Then, the model complexity of networks explodes typically exponentially as their depth increases, which makes difficult to derive optimal estimation bounds. We overcome this problem by sophisticated evaluation of model complexity using parameter rescaling techniques (see Section 5.1).

Due to its practical success, theoretical analysis for ResNet has been explored recently (Lin & Jegelka, 2018; Lu et al., 2018; Nitanda & Suzuki, 2018; Huang et al., 2018). From the viewpoint of statistical learning theory, Nitanda and Suzuki (2018) and Huang et al. (2018) investigated generalization power of ResNet from the perspective of boosting interpretation. However, they did not derive precise estimation error rates for concrete function classes. To the best of our knowledge, our theory is the first work to provide the estimation error rate of CNN classes that can accommodate the ResNet-type ones.

We import the approximation theories for FNNs, especially ones for the Barron and Hölder classes. Originally Barron (1993) considered the Barron class with a parameter $s = 1$ and an activation function σ satisfying $\sigma(z) \to 1$ as $z \to \infty$ and $\sigma(z) \to 0$ as $z \to -\infty$. Using this result, Lee et al. (2017) proved that the composition of n Barron functions with $s = 1$ can be approximated by an FNN with $n + 1$ layers. Klusowski and Barron (2018) studied its approximation theory with $s = 2$ and proved that 2-layered ReLU FNNs with M hidden units can approximate functions of this class with the order of $\tilde{O}(M^{-\frac{D+2}{2D}})$. Yarotsky (2017) proved FNNs with S non-zero parameters can approximate D variate β-Hölder continuous functions with the order of $\tilde{O}(S^{-\frac{\beta}{D}})$. Using this bound, Schmidt-Hieber (2017) proved that the estimation error of the ERM estimator is $\tilde{O}(N^{-\frac{2\beta}{2\beta+D}})$, which is minimax optimal up to logarithmic factors (see, e.g., (Tsybakov, 2008)).

3. Problem Setting

We denote the set of positive integers by $\mathbb{N}_+ := \{1, 2, \ldots\}$ and the set of positive integers less than or equal to $M \in \mathbb{N}_+$ by $[M] := \{1, \ldots, M\}$. We define $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$ for $a, b \in \mathbb{R}$.

3.1. Empirical Risk Minimization

We consider a regression task in this paper. Let X be a $[-1, 1]^D$-valued random variable with an unknown probability distribution $\mathcal{P}_X$ and ξ be an independent random noise drawn from the Gaussian distribution with an un-

Table 1. Comparison of CNN architectures. Function type: The function type CNNs can approximate. "(Block-sparse) FNNs" means any function (blocks-sparse) FNNs can realize. Channel size: the number of channles needed to approximate a β-Hölder function with accuracy ε measured by the sup norm. Sparsity: the ratio of non-zero parameters of optimal FNNs when true function is β-Hölder (N is the sample size).

	Zhou (2018)	Petersen & Voigtlaender (2018a)	Ours
CNN type	Conventional	Conventional	ResNet
Function type	Barron ($s = 2$)	FNNs	Block-sparse FNNs
Channel size	N.A.	$\tilde{O}(\varepsilon^{-\frac{D}{\beta}})$	$O(1)$
Sparsity	N.A.	$\tilde{O}(N^{-\frac{D}{2\beta+D}})$	$O(1)$

known variance σ^2 ($\sigma > 0$): $\xi \sim \mathcal{N}(0, \sigma^2)$. Let f° be an unknown deterministic function $f^\circ : [-1, 1]^D \to \mathbb{R}$ (we will characterize f° rigorously later). We define a random variable Y by $Y := f^\circ(X) + \xi$. We denote the joint distribution of (X, Y) by $\mathcal{P}$. Suppose we are given a dataset $\mathcal{D} = ((x_1, y_1), \ldots, (x_N, y_N))$ independently and identically sampled from the distribution $\mathcal{P}$, we want to estimate the true function f° from $\mathcal{D}$.

We evaluate the performance of an estimator by the squared error. For a measurable function $f : [-1, 1]^D \to \mathbb{R}$, we define the *empirical error* of f by $\hat{\mathcal{R}}_{\mathcal{D}}(f) := \frac{1}{N}\sum_{n=1}^{N}(y_n - f(x_n))^2$ and the *estimation error* by $\mathcal{R}(f) := \mathbb{E}_{X,Y}\left[(f(X) - Y)^2\right]$. Given a subset $\mathcal{F}$ of measurable functions from $[-1, 1]^D$ to $\mathbb{R}$, we consider the *clipped empirical risk minimization (ERM) estimator* $\hat{f}$ of $\mathcal{F}$ that satisfies

$$\hat{f} := \mathrm{clip}[f_{\min}] \quad \text{where } f_{\min} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \hat{\mathcal{R}}_{\mathcal{D}}(\mathrm{clip}[f]).$$

Here, clip is a clipping operator defined by $\mathrm{clip}[f] := (f \vee -\|f^\circ\|_\infty) \wedge \|f^\circ\|_\infty$. For a measurable function $f : [-1, 1]^D \to \mathbb{R}$, we define the L_2-norm (weighted by $\mathcal{P}_X$) and the sup norm of f by $\|f\|_{\mathcal{L}^2(\mathcal{P}_X)} := \left(\int_{[-1,1]^D} f^2(x)\mathrm{d}\mathcal{P}_X(x)\right)^{\frac{1}{2}}$ and $\|f\|_\infty := \sup_{x \in [-1,1]^D} |f(x)|$, respectively. Let $\mathcal{L}^2(\mathcal{P}_X)$ be the set of measurable functions f such that $\|f\|_{\mathcal{L}^2(\mathcal{P}_X)} < \infty$ with the norm $\|\cdot\|_{\mathcal{L}^2(\mathcal{P}_X)}$. The task is to estimate the *approximation* error $\inf_{f \in \mathcal{F}} \|f - f^\circ\|_\infty$ and the *estimation* error of the clipped ERM estimator: $\mathcal{R}(\hat{f}) - \mathcal{R}(f^\circ)$. Note that the estimation error is a random variable with respect the choice of the training dataset $\mathcal{D}$. By the definition of $\mathcal{R}$ and the independence of X and ξ, the estimation error equals to $\|\hat{f} - f^\circ\|^2_{\mathcal{L}^2(P_X)}$.

3.2. Convolutional Neural Networks

In this section, we define CNNs used in this paper. Let $K, C, C' \in \mathbb{N}_+$ be a filter size, input channel size, and output channel size, respectively. For a filter $w = (w_{n,j,i})_{n \in [K], j \in [C'], i \in [C]} \in \mathbb{R}^{K \times C' \times C}$, we define the *one-sided padding and stride-one convolution*[1] by w as an order-4 tensor $L^w_D = ((L^w_D)^{\beta,j}_{\alpha,i}) \in \mathbb{R}^{D \times D \times C' \times C}$ defined by

$$(L^w_D)^{\beta,j}_{\alpha,i} := \begin{cases} w_{(\alpha-\beta+1),j,i} & \text{if } 0 \leq \alpha - \beta \leq K - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here, i (resp. j) runs through 1 to C (resp. C') and α and β through 1 to D. Since we fix the input dimension D throughout the paper, we omit the subscript D and write as L^w if it is obvious from the context. We can interpret L^w as a linear mapping from $\mathbb{R}^{D \times C}$ to $\mathbb{R}^{D \times C'}$. Specifically, for $x = (x_{\alpha,i})_{\alpha,i} \in \mathbb{R}^{D \times C}$, we define $(y_{\beta,j})_{\beta,j} = L^w(x) \in \mathbb{R}^{D \times C'}$ by

$$y_{\beta,j} := \sum_{i,\alpha} (L^w)^{\beta,j}_{\alpha,i}\, x_{\alpha,i}.$$

Next, we define building blocks of CNNs: convolutional layers and fully-connected layers. Let $K, C, C' \in \mathbb{N}_+$. For a weight tensor $w \in \mathbb{R}^{K \times C' \times C}$, a bias vector $b \in \mathbb{R}^{C'}$, and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define the *convolutional layer* $\mathrm{Conv}^\sigma_{w,b} : \mathbb{R}^{D \times C} \to \mathbb{R}^{D \times C'}$ by $\mathrm{Conv}^\sigma_{w,b}(x) := \sigma(L^w(x) - \mathbf{1}_D \otimes b)$, where $\mathbf{1}_D$ is a D dimensional vector consisting of 1's, $\otimes$ is the outer product of vectors, and σ is applied in element-wise manner. Similarly, let $W \in \mathbb{R}^{C' \times DC}$, $b \in \mathbb{R}^{C'}$, and $\sigma : \mathbb{R} \to \mathbb{R}$, we define the *fully-connected layer* $\mathrm{FC}^\sigma_{W,b} : \mathbb{R}^{D \times C} \to \mathbb{R}^{C'}$ by $\mathrm{FC}^\sigma_{W,b}(a) = \sigma(W\mathrm{vec}(a) - b)$. Here, $\mathrm{vec}(\cdot)$ is the vectorization operator that flattens a matrix into a vector.

Finally, we define the ResNet-type CNN as a sequential concatenation of one convolution block, M residual blocks, and one fully-connected layer. Figure 1 is the schematic view of the CNN we adopt in this paper.

Definition 1 (Convolutional Neural Networks (CNNs))**.** *Let $M, L, C, K \in \mathbb{N}_+$, which will be the number of residual*

[1] we discuss the difference of one-sided padding and two-sided padding in Appendix H.

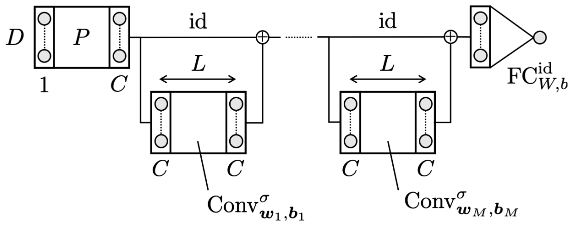
D
P
id
id
1
C
L
L
C
C
C
C
$\mathrm{Conv}^{\sigma}_{\boldsymbol{w}_1,\boldsymbol{b}_1}$
$\mathrm{Conv}^{\sigma}_{\boldsymbol{w}_M,\boldsymbol{b}_M}$
$\mathrm{FC}^{\mathrm{id}}_{W,b}$

Figure 1. ResNet-type CNN defined in Definition 1. Variables are as in Definition 1.

blocks and depth, channel size, and filter size of blocks, respectively. For $m \in [M]$ and $l \in [L]$, let $w_m^{(l)} \in \mathbb{R}^{K\times C\times C}$ and $b_m^{(l)} \in \mathbb{R}^C$ be a weight tensor and bias of the l-th layer of the m-th block in the convolution part, respectively. Finally, let $W \in \mathbb{R}^{DC\times 1}$ and $b \in \mathbb{R}$ be a weight matrix and a bias for the fully-connected layer part, respectively. For $\boldsymbol{\theta} := ((w_m^{(l)})_{m,l}, (b_m^{(l)})_{m,l}, W, b)$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define $\mathrm{CNN}^{\sigma}_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}^D$, the CNN constructed from $\boldsymbol{\theta}$, by

$$\mathrm{CNN}^{\sigma}_{\boldsymbol{\theta}} := \mathrm{FC}^{\mathrm{id}}_{W,b} \circ (\mathrm{Conv}^{\sigma}_{\boldsymbol{w}_M,\boldsymbol{b}_M} + \mathrm{id}) \circ \cdots \circ (\mathrm{Conv}^{\sigma}_{\boldsymbol{w}_1,\boldsymbol{b}_1} + \mathrm{id}) \circ P,$$

where $\mathrm{Conv}^{\sigma}_{\boldsymbol{w}_m,\boldsymbol{b}_m} := \mathrm{Conv}^{\sigma}_{w_m^{(L)},b_m^{(L)}} \circ \cdots \circ \mathrm{Conv}^{\sigma}_{w_m^{(1)},b_m^{(1)}}$, $\mathrm{id} : \mathbb{R}^{D\times C} \to \mathbb{R}^{D\times C}$ is the identity function, and $P : \mathbb{R}^D \to \mathbb{R}^{D\times C}; x \mapsto \begin{bmatrix} x & 0 & \cdots & 0 \end{bmatrix}$ is a padding operation that adds zeros to align the number of channels[2].

We say a *linear* convolutional layer or a *linear* CNN when the activation function σ is the identity function and a *ReLU* convolution layer or a *ReLU* CNN when σ is ReLU, which is defined by $\mathrm{ReLU}(x) := x \vee 0$. We borrow the term from ResNet and call $\mathrm{Conv}^{\sigma}_{\boldsymbol{w}_m,\boldsymbol{b}_m}$ $(m > 0)$ and id in the above definition the m-th *residual block* and identity mapping, respectively. We say $\boldsymbol{\theta}$ is *compatible* with (C, K) when each component of $\boldsymbol{\theta}$ satisfies the aforementioned dimension conditions.

For the number of blocks M, depth of residual blocks L, channel size C, filter size K, and norm parameters for convolution layers $B^{(\mathrm{conv})} > 0$ and for a fully-connected layer $B^{(\mathrm{fc})} > 0$, we define $\mathcal{F}^{(\mathrm{CNN})}_{M,L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$, the hypothesis class consisting of ReLU CNNs as

$$\left\{ \mathrm{CNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}} \;\middle|\; \begin{array}{l} \mathrm{CNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}} \text{ has } M \text{ residual blocks,} \\ \text{depth of each residual block is } L, \\ \boldsymbol{\theta} \text{ is compatible with } (C, K), \\ \max_{m,l} \|w_m^{(l)}\|_\infty \vee \|b_m^{(l)}\|_\infty \le B^{(\mathrm{conv})}, \\ \|W\|_\infty \vee \|b\|_\infty \le B^{(\mathrm{fc})} \end{array} \right\}.$$

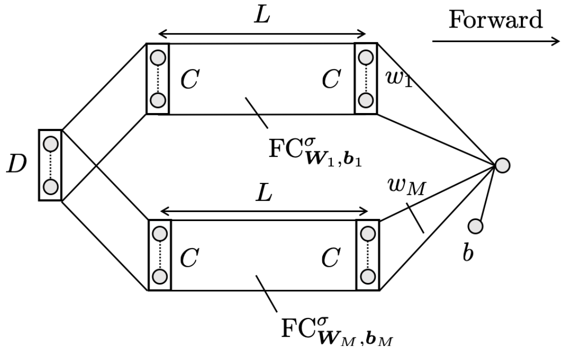
L
Forward
C
C
w_1
$\mathrm{FC}^{\sigma}_{\boldsymbol{W}_1,\boldsymbol{b}_1}$
D
L
w_M
C
C
b
$\mathrm{FC}^{\sigma}_{\boldsymbol{W}_M,\boldsymbol{b}_M}$

Figure 2. Schematic view of a block-sparse FNN. Variables are as in Definition 2.

Here, the domain of CNNs is restricted to $[-1, 1]^D$. Note that we impose norm constraints to the convolution and fully-connected part separately. We emphasize that we do not impose any sparse constraints (e.g., restricting the number of non-zero parameters in a CNN to some fixed value) on CNNs, as opposed to previous literature (Yarotsky, 2017; Schmidt-Hieber, 2017; Imaizumi & Fukumizu, 2019). We discuss differences between our CNN and the original ResNet (He et al., 2016) in Appendix I.

3.3. Block-sparse Fully-connected Neural Networks

In this section, we mathematically define FNNs we consider in this paper, in parallel with the CNN case. Our FNN, which we coin a *block-sparse* FNN, consists of M possibly dense FNNs (blocks) concatenated in parallel, followed by a single fully-connected layer. We sketch the architecture of a block-sparse FNN in Figure 2.

Definition 2 (Fully-connected Neural Networks (FNNs))**.** *Let $M, L, C \in \mathbb{N}_+$ be the number of blocks in an FNN, the depth and width of blocks, respectively. Let $W_m^{(l)} \in \mathbb{R}^{C\times C}$ and $b_m^{(l)} \in \mathbb{R}^C$ be a weight matrix and a bias of the l-th layer of the m-th block for $m \in [M]$ and $l \in [L]$, with the exception that $W_m^{(1)} \in \mathbb{R}^{C\times D}$. Let $w_m \in \mathbb{R}^C$ be a weight (sub)vector of the final fully-connected layer corresponding to the m-th block and $b \in \mathbb{R}$ be a bias for the fully-connected layer. For $\boldsymbol{\theta} = ((W_m^{(l)})_{m,l}, (b_m^{(l)})_{m,l}, (w_m)_m, b)$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define $\mathrm{FNN}^{\sigma}_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}$, the block-sparse FNN constructed from $\boldsymbol{\theta}$, by*

$$\mathrm{FNN}^{\sigma}_{\boldsymbol{\theta}} := \sum_{m=1}^{M} w_m^\top \mathrm{FC}^{\sigma}_{\boldsymbol{W}_m,\boldsymbol{b}_m}(\cdot) - b,$$

where $\mathrm{FC}^{\sigma}_{\boldsymbol{W}_m,\boldsymbol{b}_m} := \mathrm{FC}^{\sigma}_{W_m^{(L)},b_m^{(L)}} \circ \cdots \circ \mathrm{FC}^{\sigma}_{W_m^{(1)},b_m^{(1)}}$.

We say $\boldsymbol{\theta}$ is *compatible* with C when each component of $\boldsymbol{\theta}$ matches the dimension conditions determined by the width parameter C, as we did in the CNN case. When $L = 1$, a

[2] Although $\mathrm{CNN}^{\sigma}_{\boldsymbol{\theta}}$ in this definition has a fully-connected layer, we refer to a stack of convolutional layers both with or without the final fully-connect layer as a CNN in this paper.

block-sparse FNN is a 2-layered neural network with $C' := MC$ hidden units of the form $f(x) = \sum_{c=1}^{C'} b_c \sigma(a_c^\top x - t_c) - b$ where $a_c \in \mathbb{R}^D$ and $b_c, t_c, b \in \mathbb{R}$.

For the number of blocks M, depth L and width C of blocks, and norm parameters for the block part $B^{(\mathrm{bs})} > 0$ and for the final layer $B^{(\mathrm{fin})} > 0$, we define $\mathcal{F}^{(\mathrm{FNN})}_{M,L,C,B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$, the set of functions realizable by FNNs as

$$\left\{ \mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}} \;\middle|\; \begin{array}{l} \mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}} \text{ has } M \text{ blocks,} \\ \text{depth of each block is } L, \\ \boldsymbol{\theta} \text{ is compatible with } C, \\ \max_{m,l} \|W^{(l)}_m\|_\infty \vee \|b^{(l)}_m\|_\infty \leq B^{(\mathrm{bs})}, \\ \max_m \|w_m\|_\infty \vee |b| \leq B^{(\mathrm{fin})}. \end{array} \right\},$$

where the domain is again restricted to $[-1, 1]^D$.

4. Main Theorems

With the preparation in previous sections, we state our main results of this paper. We only describe statements of theorems and corollaries in the main article. All complete proofs are deferred to the supplemental material.

4.1. Approximation

Our first theorem claims that any block-sparse FNN with M blocks is realizable by a ResNet-type CNN with fixed-sized channels and filters by adding $O(M)$ parameters.

Theorem 1. *Let $M, L, C \in \mathbb{N}_+$, $K \in \{2, \ldots D\}$ and $L_0 := \left\lceil \frac{D-1}{K-1} \right\rceil$. Then, there exist $L' \leq L + L_0$, $C' \leq 4C$, and $K' \leq K$ such that, for any $B^{(\mathrm{bs})}, B^{(\mathrm{fin})} > 0$, any FNN in $\mathcal{F}^{(\mathrm{FNN})}_{M,L,C,B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$ can be realized by a CNN in $\mathcal{F}^{(\mathrm{CNN})}_{M,L',C',K',B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$. Here, $B^{(\mathrm{conv})} = B^{(\mathrm{bs})}$ and $B^{(\mathrm{fc})} = B^{(\mathrm{fin})}(1 \vee (B^{(bs)})^{-1})$.*

In particular, if we can approximate a function with a block-sparse FNN with $O(M)$ parameters, we can approximate the function with a ResNet-type CNN at the same rate, too. By the definition of $\mathcal{F}^{(\mathrm{CNN})}_{M,L',C',K',B^{(\mathrm{conv})}}$, the CNN emulating the block-sparse FNN is dense and does not have sparse structures in general.

4.2. Estimation

Our second theorem bounds the estimation error of the clipped ERM estimator. We denote $\mathcal{F}^{(\mathrm{FNN})} = \mathcal{F}^{(\mathrm{FNN})}_{M,L,C,B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$ and $\mathcal{F}^{(\mathrm{CNN})} = \mathcal{F}^{(\mathrm{CNN})}_{M,L',C',K',B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$ in short.

Theorem 2. *Let $f^\circ : \mathbb{R}^D \to \mathbb{R}$ be a measurable function and $B^{(\mathrm{bs})}, B^{(\mathrm{fin})} > 0$. Let M, L, C, K, and L_0 as in Theorem 1. Suppose $L', C', K', B^{(\mathrm{conv})}$ and $B^{(\mathrm{fc})}$ satisfy $\mathcal{F}^{(\mathrm{FNN})} \subset \mathcal{F}^{(\mathrm{CNN})}$ (their existence is ensured by Theorem 1). Suppose that the covering nubmer of $\mathcal{F}^{(\mathrm{CNN})}$ is larger than 2. Then, the clipped ERM estimator $\hat{f}$ of $\mathcal{F} := \{\mathrm{clip}[f] \mid f \in \mathcal{F}^{(\mathrm{CNN})}\}$ satisfies*

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}} \|\hat{f} - f^\circ\|^2_{\mathcal{L}^2(\mathcal{P}_X)} \\ &\leq C_0 \left(\inf_f \|f - f^\circ\|^2_\infty + \frac{\tilde{F}^2}{N} \Lambda_2 \log(2\Lambda_1 BN) \right). \end{aligned} \quad (1)$$

Here, f ranges over $\mathcal{F}^{(\mathrm{FNN})}$, $C_0 > 0$ is a universal constant, $\tilde{F} := \frac{\|f^\circ\|_\infty}{\sigma} \vee \frac{1}{2}$, and $B := B^{(\mathrm{conv})} \vee B^{(\mathrm{fc})}$. $\Lambda_1 = \Lambda_1(\mathcal{F}^{(\mathrm{CNN})})$ and $\Lambda_2 = \Lambda_2(\mathcal{F}^{(\mathrm{CNN})})$ are defined by

$$\Lambda_1 := (2M + 3)C'D(1 \vee B^{(\mathrm{fc})})(1 \vee B^{(\mathrm{conv})})\varrho\varrho^+,$$
$$\Lambda_2 := ML' \left(C'^2 K' + C' \right) + C'D + 1,$$

where $\varrho := (1 + \rho)^M$, $\varrho^+ := 1 + ML'\rho^+$, $\rho := (C'K'B^{(\mathrm{conv})})^{L'}$, and $\rho^+ := (1 \vee C'K'B^{(\mathrm{conv})})^{L'}$.

The first term of (1) is the approximation error achieved by $\mathcal{F}^{(\mathrm{FNN})}$. On the other hand, the second term of (1) represents the model complexity of $\mathcal{F}^{(\mathrm{CNN})}$ since Λ_1 and Λ_2 are determined by the architectural parameters of $\mathcal{F}^{(\mathrm{CNN})}$ — Λ_1 corresponds to the Lipschitz constant of a function realized by a CNN and Λ_2 is the number of parameters, including zeros, of a CNN. There is a trade-off between these two terms. Using appropriately chosen M to balance them, we can evaluate the order of estimation error with respect to the sample size N.

Corollary 1. *Under the same assumptions as Theorem 2, suppose further $\log \Lambda_1 B = \tilde{O}(1)$ as a function of M. If $\inf_{f \in \mathcal{F}^{(\mathrm{FNN})}} \|f - f^\circ\|^2_\infty = \tilde{O}(M^{-\gamma_1})$ and $\Lambda_2 = \tilde{O}(M^{\gamma_2})$ for some constants $\gamma_1, \gamma_2 > 0$ independent of M, then, the clipped ERM estimator $\hat{f}$ of $\mathcal{F}$ achieves the estimation error $\|f^\circ - \hat{f}\|^2_{\mathcal{L}_2(\mathcal{P}_X)} = \tilde{O}_P(N^{-\frac{2\gamma_1}{2\gamma_1 + \gamma_2}})$.*

5. Application

5.1. Barron Class

The Barron class is an example of the function class that can be approximated by block-sparse FNNs. We employ the definition of Barron functions used in (Klusowski & Barron, 2018).

Definition 3 (Barron class)**.** *We call a measurable function $f^\circ : [-1, 1]^D \to \mathbb{R}$ a Barron function of a parameter $s > 0$ if f° admits the Fourier representation (i.e., $f^\circ(x) = \check{\mathcal{F}}\mathcal{F}[f^\circ]$) and $\int_{\mathbb{R}^D} \|w\|_2^s |\mathcal{F}[f^\circ](w)| \, \mathrm{d}w < \infty$. Here, $\mathcal{F}$ and $\check{\mathcal{F}}$ are the Fourier and inverse Fourier transformation, respectively.*

Klusowski and Barron (2018) studied approximation of the Barron function f° with the parameter $s = 2$ by a linear combination of M ridge functions (i.e., a 2-layered ReLU FNN). Specifically, they showed that there exists a function f_M of the form

$$f_M := f^\circ(0) + \nabla f^{\circ\top}(0)x + \frac{1}{M}\sum_{m=1}^{M} b_m(a_m^\top x - t_m)_+$$

with $|b_m| \leq 1$, $\|a_m\|_1 = 1$, and $|t_m| \leq 1$, such that $\|f^\circ - f_M\|_\infty = \tilde{O}(M^{-\left(\frac{1}{2}+\frac{1}{D}\right)})$. Using this approximator f_M, we can derive the same approximation order using CNNs by applying Theorem 1 with $L = 1$ and $C = 1$.

Corollary 2. *Let $f^\circ : [-1,1]^D \to \mathbb{R}$ be a Barron function with the parameter $s = 2$ such that $f^\circ(0) = 0$ and $\nabla f^\circ(0) = \mathbf{0}_D$. Then, for any $K \in \{2, \ldots, D\}$, there exists a CNN $f^{(\mathrm{CNN})}$ with M residual blocks, each of which has depth $O(1)$ and at most 4 channels, and whose filter size is at most K, such that $\|f^\circ - f^{(\mathrm{CNN})}\|_\infty = \tilde{O}(M^{-\left(\frac{1}{2}+\frac{1}{D}\right)})$.*

Note that this rate is same as the one obtained for FNNs (Klusowski & Barron, 2018).

We have one design choice when we apply Corollary 1 in order to derive the estimation error: how to set $B^{(\mathrm{bs})}$ and $B^{(\mathrm{fin})}$? Looking at the definition of f_M, a naive choice would be $B^{(\mathrm{bs})} := 1$ and $B^{(\mathrm{fin})} := M^{-1}$. However, this cannot satisfy the assumption on Λ_1 of Corollary 1, due to the term $\varrho = (1+\rho)^M$. We want the logarithm of Λ_1 to be $\tilde{O}(1)$ as a function of M. In order to do that, we change the *relative scale* between parameters in the block-sparse part and the fully-connected part using the homogeneous property of the ReLU function: $\mathrm{ReLU}(ax) = a\mathrm{ReLU}(x)$ for $a > 0$. The rescaling operation enables us to choose $B^{(\mathrm{bs})} := M^{-1}$ and $B^{(\mathrm{fin})} = 1$ to meet the assumption of Corollary 1. By setting $\gamma_1 = \frac{1}{2} + \frac{1}{D}$ and $\gamma_2 = 1$, we obtain the desired estimation error.

Corollary 3. *Let $f^\circ : [-1,1]^D \to \mathbb{R}$ be a Barron function with the parameter $s = 2$ such that $f^\circ(0) = 0$ and $\nabla f^\circ(0) = \mathbf{0}_D$. Let $K \in \{2, \ldots, D\}$. There exist the number of residual blocks $M = O(N^{\frac{D}{2+2D}})$, depth of each residual block $L = O(1)$, channel size $C = O(1)$, and norm bounds $B^{(\mathrm{conv})}, B^{(\mathrm{fc})} > 0$ such that for sufficiently large N, the clipped ERM estimator $\hat{f}$ of $\{\mathrm{clip}[f] \mid f \in \mathcal{F}^{(\mathrm{CNN})}_{M,L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fc})}}\}$ achieves the estimation error $\|f^\circ - \hat{f}\|^2_{\mathcal{L}_2(\mathcal{P}_X)} = \tilde{O}_P(N^{-\frac{D+2}{2(D+1)}})$.*

5.2. Hölder Class

We next consider the approximation and error rates of CNNs when the true function f° is a Hölder function.

Definition 4 (Hölder class)**.** *Let $\beta > 0$. A function $f^\circ : [-1,1]^D \to \mathbb{R}$ is called a β-Hölder function if*

$$\|f^\circ\|_\beta := \sum_{0 \leq |\alpha| < \lfloor \beta \rfloor} \|\partial^\alpha f^\circ\|_\infty + \sum_{|\alpha| = \lfloor \beta \rfloor} \sup_{x \neq y} \frac{|\partial^\alpha f^\circ(x) - \partial^\alpha f^\circ(y)|}{|x-y|^{\beta - \lfloor \beta \rfloor}} < \infty.$$

Here, $\alpha = (\alpha_1, \ldots, \alpha_D)$ is a multi-index. That is, $\partial^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_D^{\alpha_D}}$ and $|\alpha| := \sum_{d=1}^D \alpha_d$.

Yarotsky (2017) showed that FNNs with S non-zero parameters can approximate any D variate β-Hölder function with the order of $\tilde{O}(S^{-\frac{\beta}{D}})$. Schmidt-Hieber (2017) also proved a similar statement using a different construction method. They only specified the width[3], depth, and non-zero parameter counts of the approximating FNN and did not write in detail how non-zero parameters are distributed in the statements explicitly (see Theorem 1 of (Yarotsky, 2017) and Theorem 5 of (Schmidt-Hieber, 2017)). However, if we carefully look at their proofs, we find that we can transform the FNNs they constructed into block-sparse ones (see Lemma 7 of the supplemental material). Therefore, we can apply Theorem 1 to these FNNs. To meet the assumption of Corollary 1, we again rescale the parameters of the FNNs, as we did in the Barron-class case, so that $\log \Lambda_1 = \tilde{O}(1)$. We can derive the approximation and estimation errors by setting $\gamma_1 = \frac{\beta}{D}$ and $\gamma_2 = 1$.

Corollary 4. *Let $\beta > 0$ and $f^\circ : [-1,1]^D \to \mathbb{R}$ be a β-Hölder function. Then, for any $K \in \{2, \ldots, D\}$, there exists a CNN $f^{(\mathrm{CNN})}$ with $O(M)$ residual blocks, each of which has depth $O(\log M)$ and $O(1)$ channels, and whose filter size is at most K, such that $\|f^\circ - f^{(\mathrm{CNN})}\|_\infty = \tilde{O}(M^{-\frac{\beta}{D}})$.*

Corollary 5. *Let $\beta > 0$ and $f^\circ : [-1,1]^D \to \mathbb{R}$ be a β-Hölder function. For any $K \in \{2, \ldots, D\}$, there exist the number of residual blocks $M = O(N^{\frac{D}{2\beta+D}})$, depth of each residual block $L = O(\log N)$, channel size $C = O(1)$, and norm bounds $B^{(\mathrm{conv})}, B^{(\mathrm{fc})} > 0$ such that for sufficiently large N, the clipped ERM estimator $\hat{f}$ of $\{\mathrm{clip}[f] \mid f \in \mathcal{F}^{(\mathrm{CNN})}_{M,L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fc})}}\}$ achieves the estimation error $\|f^\circ - \hat{f}\|^2_{\mathcal{L}_2(\mathcal{P}_X)} = \tilde{O}_P(N^{-\frac{2\beta}{2\beta+D}})$.*

Since the estimation error rate of the β-Hölder class is $O_P(N^{-\frac{2\beta}{2\beta+D}})$ (see, e.g., (Tsybakov, 2008)), Corollary 5 implies that our CNN can achieve the minimax optimal rate up to logarithmic factors even though it can be dense and its width D, channel size C, and filter size K are constant with respect to the sample size N.

[3] Yarotsky (2017) didn't specified the width of FNNs.

6. Optimal CNNs with Constant-depth Blocks

In the previous section, we proved the optimality of dense and narrow ResNet-type CNNs for the Hölder class. However, the constructed CNN can have residual blocks whose depth is as large as $O(\log N)$. Such an architecture is different from practically successful ResNets because they usually have relatively shallow (e.g., 2- or 3-layered) networks as residual blocks. We hypothesize that the essence of the problem resides in the difference of scales between identity connections and residual blocks. Therefore, we consider another type of CNNs that admits scaling schemes of intermediate signals in order to overcome this problem. Among others, we consider the simplest scaling method, which zeros out some channels in identity mappings.

Definition 5 (Masked CNNs)**.** *Let $M, L, C, K \in \mathbb{N}_+$. Let $w_m^{(l)} \in \mathbb{R}^{K\times C\times C}$, $b_m^{(l)} \in \mathbb{R}^C$, $W \in \mathbb{R}^{DC\times 1}$ and $b \in \mathbb{R}$ be parameters of CNNs for $m \in [M]$ and $l \in [L]$. Let $z_m = (z_{m,1}, \ldots, z_{m,C}) \in \{0,1\}^C$ be a mask for the m-th identity mapping. For $\boldsymbol{\theta} := ((w_m^{(l)})_{m,l}, (b_m^{(l)})_{m,l}, W, b, (z_m)_m)$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define $\mathrm{mCNN}_{\boldsymbol{\theta}}^{\sigma} : \mathbb{R}^D \to \mathbb{R}^D$, the masked CNN constructed from $\boldsymbol{\theta}$, by*

$$\mathrm{mCNN}_{\boldsymbol{\theta}}^{\sigma} := \mathrm{FC}_{W,b}^{\mathrm{id}} \circ (\mathrm{Conv}_{\boldsymbol{w}_M,\boldsymbol{b}_M}^{\sigma} + J_M) \circ \cdots \circ (\mathrm{Conv}_{\boldsymbol{w}_1,\boldsymbol{b}_1}^{\sigma} + J_1) \circ P,$$

where $J_m : \mathbb{R}^{D\times C} \to \mathbb{R}^{D\times C}$ is a channel wise mask operation defined by $[x_1 \ \cdots \ x_C] \mapsto [z_{m,1}x_1 \ \cdots \ z_{m,C}x_C]$.

By definition, plain ResNet-type CNNs in Definition 1 are a special case of masked CNNs. Note that we do not restrict the number of non-zero mask elements. Therefore, although masks take discrete values, we can obtain approximated ERM estimators via sparse optimization techniques. We say $\boldsymbol{\theta}$ is compatible with (C, K) when $\boldsymbol{\theta}$ satisfies the dimension conditions as we did in Definition 1. We define $\mathcal{G}_{M,L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$ by

$$\left\{ \mathrm{mCNN}_{\boldsymbol{\theta}}^{\mathrm{ReLU}} \;\middle|\; \begin{array}{l} \mathrm{mCNN}_{\boldsymbol{\theta}}^{\mathrm{ReLU}} \text{ has } M \text{ residual blocks,} \\ \text{depth of each residual block is } L, \\ \boldsymbol{\theta} \text{ is compatible with } (C, K), \\ \max_{m,l} \|w_m^{(l)}\|_\infty \vee \|b_m^{(l)}\|_\infty \le B^{(\mathrm{conv})}, \\ \|W\|_\infty \vee \|b\|_\infty \le B^{(\mathrm{fc})} \end{array} \right\}.$$

In the above definition, we treat the mask pattern $z = (z_m)_m$ as learnable parameters. We can also treat z as fixed during training and search for best z as architecture search. The following theorems show that masked CNNs can approximate and estimate any Hölder function optimally even if the depth of residual blocks is specified *a priori*. We treat L as a constant against M in the theorems.

Theorem 3. *Let $f^\circ : [-1,1]^D \to \mathbb{R}$ be a β-Hölder function. For any $K \in \{2, \ldots, D\}$ and $L \in \mathbb{N}_+$, there exists a CNN $f^{(\mathrm{CNN})}$ with $O(M \log M)$ residual blocks, each of which has depth L and $O(1)$ channels, and whose filter size is at most K, such that $\|f^\circ - f^{(\mathrm{CNN})}\|_\infty = \tilde{O}(M^{-\frac{\beta}{D}})$.*

Theorem 4. *Let $f^\circ : [-1,1]^D \to \mathbb{R}$ be a β-Hölder function. For any $K \in \{2, \ldots, D\}$ and $L \in \mathbb{N}_+$, there exist the number of residual blocks $\tilde{M} = O(N^{\frac{D}{2\beta+D}} \log N)$, channel size $C = O(1)$, and norm bounds $B^{(\mathrm{conv})}, B^{(\mathrm{fc})} > 0$ such that for sufficiently large N, the clipped ERM estimator $\hat{f}$ of $\{\mathrm{clip}[f] \mid f \in \mathcal{G}_{\tilde{M},L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fc})}}\}$ achieves the estimation error $\|f^\circ - \hat{f}\|_{\mathcal{L}_2(\mathcal{P}_X)}^2 = \tilde{O}_P(N^{-\frac{2\beta}{2\beta+D}})$.*

7. Conclusion

In this paper, we established new approximation and statistical learning theories for CNNs by utilizing the ResNet-type architecture of CNNs and the block-sparse structure of FNNs. We proved that any block-sparse FNN with M blocks is realizable by a CNN that has $O(M)$ additional parameters. Then, we derived the approximation and estimation error rates for CNNs from those for block-sparse FNNs. Our theory is general in a sense that it does not depend on a specific function class, as long as we can approximate it with block-sparse FNNs. Using this theory, we derived approximation and error rates for the Barron and Hölder classes in almost the same manner and showed that the estimation error of CNNs is same as that of FNNs, even if CNNs are dense and have constant channel size, filter size, and width with respect to the sample size. We can additionally make the depth of residual blocks constant if we allowed identity mappings to have scaling schemes. The key techniques were careful evaluations of the Lipschitz constant and non-trivial weight parameter rescaling of NNs.

One of the interesting open questions is the role of the weight rescaling. We critically use the homogeneous property of the ReLU to change the relative scale between the block-sparse and fully-connected part, if it were not for this property, the estimation error rate would be worse. The general theory for rescaling, not restricted to the Barron nor Hölder classes would be beneficial for deeper understanding of the relationship between the approximation and estimation capabilities of FNNs and CNNs.

Another question is when the approximation and estimation error rates of CNNs can *exceed* that of FNNs. We can derive the same rates as FNNs essentially because we can realize block-sparse FNNs using CNNs that have the same order of parameters (see Theorem 1). If we can find some special structures of FNNs – like repetition, then, the CNNs might need fewer parameters and can achieve better estimation error rate. Note that there is no hope for enhancement for the Hölder case since the estimation rate using FNNs is already minimax optimal (up to logarithmic factors). It is left for future research which function classes and constraints of FNNs, like block-sparseness, we should choose.

Acknowledgements

We thank Kohei Hayashi for improving on the draft, Wei Lu for commenting on the preprint version of the paper and pointing out its errata, and anonymous reviewers for fruitful discussion and positive feedback and comments. TS was partially supported by MEXT Kakenhi (26280009, 15H05707, 18K19793 and 18H03201), Japan Digital Design and JSTCREST.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263. PMLR, 2018.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

Chen, M., Pennington, J., and Schoenholz, S. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 873–882. PMLR, 2018.

Cohen, T. S., Geiger, M., Khler, J., and Welling, M. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Fan, F., Wang, D., and Wang, G. Universal approximation by a slim network with sparse shortcut connections. *arXiv preprint arXiv:1811.09003*, 2018.

Giné, E. and Nickl, R. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Huang, F., Ash, J., Langford, J., and Schapire, R. Learning deep ResNet blocks sequentially using boosting theory. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2058–2067. PMLR, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269. IEEE, 2017.

Imaizumi, M. and Fukumizu, K. Deep neural networks learn non-smooth functions effectively. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 869–878. PMLR, 2019.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015.

Kainen, P. C., Kůrková, V., and Sanguineti, M. Approximating multivariable functions by feedforward neural nets. In *Handbook on Neural Information Processing*, pp. 143–181. Springer, 2013.

Klusowski, J. M. and Barron, A. R. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ_1 and ℓ_0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

Lee, H., Ge, R., Ma, T., Risteski, A., and Arora, S. On the ability of neural nets to express distributions. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1271–1296. PMLR, 2017.

Lin, H. and Jegelka, S. ResNet with one-neuron hidden layers is a universal approximator. In *Advances in Neural Information Processing Systems 31*, pp. 6169–6178. Curran Associates, Inc., 2018.

Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3276–3285. PMLR, 2018.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: a view from the width. In *Advances in Neural Information Processing Systems 30*, pp. 6231–6239. Curran Associates, Inc., 2017.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Nitanda, A. and Suzuki, T. Functional gradient boosting based on residual network perception. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3819–3828. PMLR, 2018.

Perekrestenko, D., Grohs, P., Elbrächter, D., and Bölcskei, H. The universal approximation power of finite-width deep ReLU networks. *arXiv preprint arXiv:1806.01528*, 2018.

Petersen, P. and Voigtlaender, F. Equivalence of approximation by convolutional neural networks and fully-connected networks. *arXiv preprint arXiv:1809.00973*, 2018a.

Petersen, P. and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018b.

Pinkus, A. Density in approximation theory. *Surveys in Approximation Theory*, 1:1–45, 2005.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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

Suzuki, T. Fast generalization error bound of deep learning from a kernel perspective. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1397–1406. PMLR, 2018.

Suzuki, T. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Yarotsky, D. Universal approximations of invariant maps by neural networks. *arXiv preprint arXiv:1804.10306*, 2018.

Zhou, D.-X. Universality of deep convolutional neural networks. *arXiv preprint arXiv:1805.10769*, 2018.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931, 2015.

Zhou, P. and Feng, J. Understanding generalization and optimization performance of deep CNNs. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5960–5969. PMLR, 2018.