Approximation and Non-parametric Estimation of ResNet-type Convolutional Neural Networks (Supplemental Material)

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

Abstract

In this supplemental material, we give proofs of theorems and corollaries in the main article. We prove them in more general form. Specifically, we allow CNNs to have residual blocks with different depth and each residual block to have varying numbers of channels and filter sizes. Similarly, FNNs can have blocks with different depth, and the width of a block can be non-constant.

A. Notation

For tensor a, we define the positive part of a by $a_+ := a \vee 0$ where the maximum operation is performed in element-wise manner. Similarly the negative part of a is defined as $a_- := -a \vee 0$. Note that $a = a_+ - a_-$ holds for any tensor a. For normed spaces $(V, \|\cdot\|_V), (W, \|\cdot\|_W)$ and a linear operator $T : V \to W$ we denote the operator norm of T by $\|T\|_{\mathrm{op}} := \sup_{\|v\|_V=1} \|Tv\|_W$. For a sequence $\boldsymbol{w} = (w^{(1)}, \ldots, w^{(L)})$ and $l \leq l'$, we denote its subsequence from the l-th to l'-th elements by $\boldsymbol{w}[l : l'] := (w^{(l)}, \ldots, w^{(l')})$.

B. Definitions

We define general types of ResNet-type CNNs and block-sparse FNNs.

Definition 6 (Convolutional Neural Networks (CNNs))**.** *Let $M \in \mathbb{N}_+$ and $L_m \in \mathbb{N}_+$, which will be the number of residual blocks and the depth of m-th block, respectively. Let $C_m^{(l)}, K_m^{(l)}$ be the channel size and filter size of the l-th layer of the m-th block for $m \in [M]$ and $l \in [L_m]$. We assume $C_1^{(L_1)} = \cdots = C_M^{(L_M)}$ and denote it by $C^{(0)}$. Let $w_m^{(l)} \in \mathbb{R}^{K_m^{(l)} \times C_m^{(l)} \times C_m^{(l-1)}}$ and $b_m^{(l)} \in \mathbb{R}$ be the weight tensors and biases of l-th layer of the m-th block in the convolution part, respectively. Here $C_m^{(0)}$ is defined as $C^{(0)}$. Finally, let $W \in \mathbb{R}^{D \times C_0^{(L_0)}}$ and $b \in \mathbb{R}$ be the weight matrix and the 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 $\mathrm{CNN}_{\boldsymbol{\theta}}^{\sigma} : \mathbb{R}^D \to \mathbb{R}^D$, the CNN constructed from $\boldsymbol{\theta}$, by*

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

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

Definition 7 (Fully-connected Neural Networks (FNNs))**.** *Let $M \in \mathbb{N}_+$ be the number of blocks in an FNN. Let $\boldsymbol{D}_m = (D_m^{(1)}, \ldots, D_m^{(L_m)}) \in \mathbb{N}_+^{L_m}$ be the sequence of intermediate dimensions of the m-th block, where $L_m \in \mathbb{N}_+$ is the depth of the m-th block for $m \in [M]$. Let $W_m^{(l)} \in \mathbb{R}^{D_m^{(l)} \times D_m^{(l-1)}}$ and $b_m^{(l)} \in \mathbb{R}^{D_m^{(l)}}$ be the weight matrix and the bias of the l-th layer*

[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.

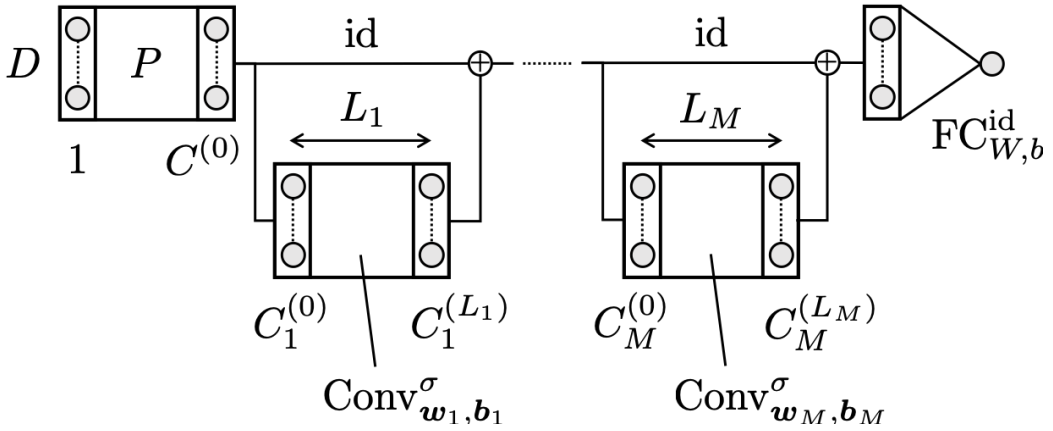

Figure 3. ResNet-type CNN defined in Definition 6. Variables are as in Definition 6.

of m-th block (with the convention $D_m^{(0)} = D$). Let $w_m \in \mathbb{R}^{D_m^{(L_m)}}$ be the weight (sub)vector of the final fully-connected layer corresponding to the m-th block and $b \in \mathbb{R}$ be the bias for the last 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 $\mathrm{FNN}_{\boldsymbol{\theta}}^{\sigma} : \mathbb{R}^D \to \mathbb{R}$, the block-sparse FNN constructed from $\boldsymbol{\theta}$, by

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

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

Figure 3 shows the schematic view of a ResNet-type CNNs defined in Definition 6 and Figure 4 shows that of Definition 7. Definition 6 is reduced to Definition 1 by setting $L_m = L$, $\boldsymbol{C} = (C)_{m,l}$ and $\boldsymbol{K} = (K)_{m,l}$. Similarly, Definition 2 is a special case of Definition 7 where $L_m = L$ and $\boldsymbol{D} = (C)_{m,l}$. Correspondingly, we denote the set of functions realizable by CNNs and FNNs by $\mathcal{F}^{(\mathrm{CNN})}_{\boldsymbol{C},\boldsymbol{K},B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$ and $\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$, respectively [1].

C. Proof of Theorem 1

We restate Theorem 1 in more general form. Note that Theorem 1 is a special case of Theorem 5 where width, depth, channel sizes and filter sizes are same among blocks.

Theorem 5. *Let $M \in \mathbb{N}_+$, $K \in \{2, \dots D\}$, and $L_0 := \left\lceil \frac{D-1}{K-1} \right\rceil$. Let $L_m, D_m^{(l)} \in \mathbb{N}_+$ and $\boldsymbol{D} = (D_m^{(l)})_{m,l}$ for $m \in [M]$ and $l \in [L_m]$. Then, there exist $L'_m \in \mathbb{N}_+$, $\boldsymbol{C} = (C_m^{(l)})_{m,l}$, and $\boldsymbol{K} = (K_m^{(l)})_{m,l}$ $(m \in [M], l \in [L'_m])$ satisfying the following properties:*

1. $L'_m \le L_m + L_0$ *$(\forall m \in [M])$,*
2. $\max_{l \in [L'_m]} C_m^{(l)} \le 4 \max_{l \in [L_m]} D_m^{(l)}$ *$(\forall m \in [M])$, and*
3. $\max_{l \in [L'_m]} K_m^{(l)} \le K$ *$(\forall m \in [M], \forall l \in [L'_m])$*

such that for any $B^{(\mathrm{bs})}, B^{(\mathrm{fin})} > 0$, any FNN in $\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$ can be realized by a CNN in $\mathcal{F}^{(\mathrm{CNN})}_{\boldsymbol{C},\boldsymbol{K},B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$. Here, $B^{(\mathrm{conv})} = B^{(\mathrm{bs})}$ and $B^{(\mathrm{fc})} = B^{(\mathrm{fin})}(1 \vee (B^{(\mathrm{bs})})^{-1})$. Further, if $L_1 = \cdots = L_M$, then we can choose L'_m to be a same value.

Remark 1. *For $K \le K'$, we can embed $\mathbb{R}^K$ into $\mathbb{R}^{K'}$ by inserting zeros: $w = (w_1, \dots, w_K) \mapsto w' = (w_1, \dots, w_K, 0, \dots, 0)$. It is easy to show $L^w = L^{w'}$. Using this equality, we can expand a size-K filter to size-K'.*

[1] Note that information of M and L_m are included in $\boldsymbol{C}$, $\boldsymbol{K}$, and $\boldsymbol{D}$. Therefore, we do not have to put them as subscripts

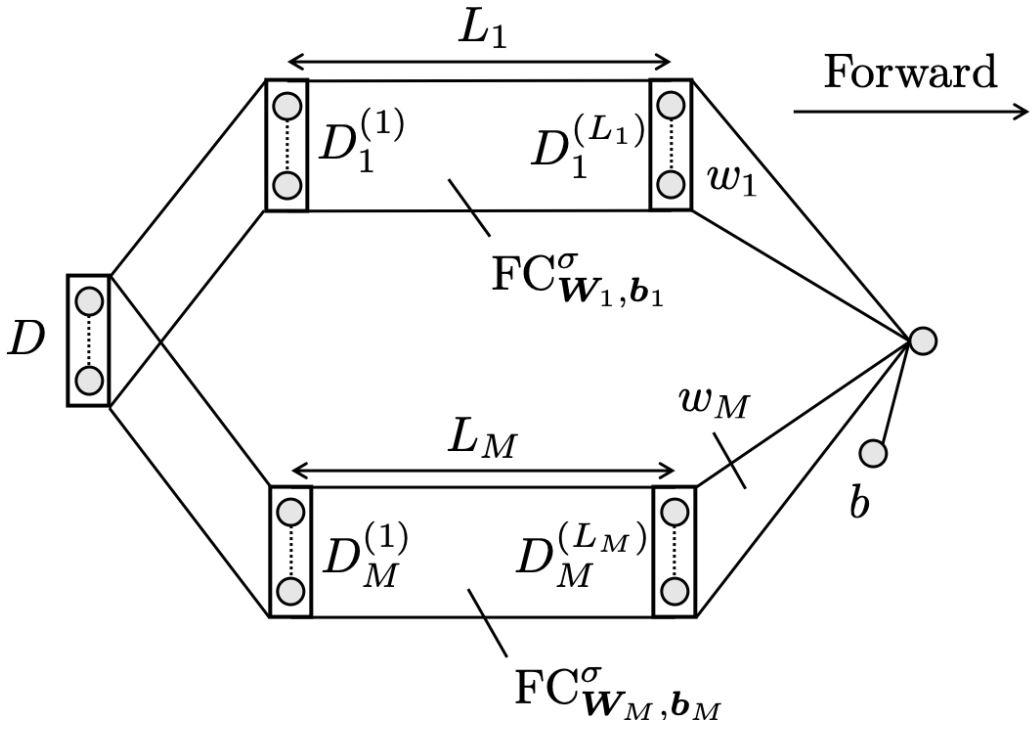

Figure 4. Schematic view of a block-sparse FNN. Variables are as in Definition 7.

Furthermore, we can arbitrary increase the number of output channels of a convolution layer by adding filters consisting of zeros. Therefore, although properties 2 and 3 allow $C_m^{(l)}$ and $K_m^{(l)}$ to be different values, we can choose $C_m^{(l)}$ and $K_m^{(l)}$ so that inequalities in property 2. and 3. are actually equals by adding filters consisting of zeros. In particular, when $D_m^{(l)}$'s are same value, we can choose $C_m^{(l)}$ to be same.

C.1. Proof Overview

For $f^{(\mathrm{FNN})} \in \mathcal{F}^{(\mathrm{FNN})}$, we realize a CNN $f^{(\mathrm{CNN})}$ using M residual blocks by "serializing" blocks in the FNN and converting them into convolution layers.

First we multiply the channel size by three using the first padding operation. We will use the first channel for storing the original input signal for feeding to downstream blocks and the second and third ones for accumulating properly scaled outputs of each blocks, that is, $\sum_{m=1}^{m'} w_m^\top \mathrm{FC}^{\mathrm{ReLU}}_{\boldsymbol{W}_m,\boldsymbol{b}_m}(x)$ where w_m is the weight of the final fully-connected layer corresponding to the m-th block.

For $m = 1, \ldots, M$, we create the m-th residual block from the m-th block of $f^{(\mathrm{FNN})}$. First, we show that for any $a \in \mathbb{R}^D$ and $t \in \mathbb{R}$, there exists L_0-layered 4-channel ReLU CNN with $O(D)$ parameters whose first output coordinate equals to a ridge function $x \mapsto (a^\top x - t)_+$ (Lemma 1 and Lemma 2). Since the first layer of m-th block is concatenation of C hinge functions, it is realizable by a $4C$-channel ReLU CNN with L_0-layers.

For the l-th layer of the m-th block ($m \in [M], l = 2, \ldots, L'_m$), we prepare C size-1 filters made from the weight parameters of the corresponding layer of the FNN. Observing that the convolution operation with size-1 filter is equivalent to a dimension-wise affine transformation, the first coordinate of the output of l-th layer of the CNN is inductively same as that of the m-th block of the FNN. After computing the m-th block FNN using convolutions, we add its output to the accumulating channel in the identity mapping.

Finally, we pick the first coordinate of the accumulating channel and subtract the bias term using the final affine transformation.

C.2. Decomposition of Affine Transformation

The following lemma shows that any affine transformation is realizable with a $\left\lceil \frac{D-1}{K-1} \right\rceil$-layered linear conventional CNN (without the final fully-connect layer).

Lemma 1. *Let $a \in \mathbb{R}^D$, $t \in \mathbb{R}$, $K \in \{2, \dots, D-1\}$, and $L_0 := \left\lceil \frac{D-1}{K-1} \right\rceil$. Then, there exists*

$$w^{(l)} \in \begin{cases} \mathbb{R}^{K \times 2 \times 1} & \text{(for } l = 1\text{)} \\ \mathbb{R}^{K \times 2 \times 2} & \text{(for } l = 2, \dots, L_0 - 1\text{)} \\ \mathbb{R}^{K \times 1 \times 2} & \text{(for } l = L_0\text{)} \end{cases}$$

and $b \in \mathbb{R}$ such that

1. $\max_{l \in [L_o]} \|w_m\|_\infty = \|a\|_\infty$, $\max_{l \in [L_0]} \|b^{(l)}\|_\infty = |t|$, *and*
2. $\mathrm{Conv}^{\mathrm{id}}_{\boldsymbol{w},\boldsymbol{b}} : \mathbb{R}^D \to \mathbb{R}^D$ *satisfies* $\mathrm{Conv}^{\mathrm{id}}_{\boldsymbol{w},\boldsymbol{b}}(x) = a^\top x - t$ *for any* $x \in [-1, 1]^D$.

Proof. First, observe that the convolutional layer constructed from $u = \begin{bmatrix} u_1 & \dots & u_K \end{bmatrix}^\top \in \mathbb{R}^{K \times 1 \times 1}$ takes the inner product with the first K elements of the input signal: $L^u(x) = \sum_{k=1}^K u_k x_k$. In particular, $u = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}^\top \in \mathbb{R}^{K \times 1 \times 1}$ works as the "left-translation" by $K - 1$. Therefore, we should define $\boldsymbol{w}$ so that it takes the inner product with the K left-most elements in the first channel and shift the input signal by $K - 1$ with the second channel. Specifically, we define $\boldsymbol{w} = (w^{(1)}, \dots, w^{(L_0)})$ by

$$(w^{(1)})_{:,1,:} = \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix}, \qquad (w^{(1)})_{:,2,:} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

$$(w^{(l)})_{:,1,:} = \begin{bmatrix} 0 & a_{(l-1)K+1} \\ \vdots & \vdots \\ 0 & a_{lK} \end{bmatrix}, \quad (w^{(l)})_{:,2,:} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 0 \end{bmatrix},$$

$$(w^{(L_0)})_{:,1,:} = \begin{bmatrix} 0 & a_{(L_0-1)K+1} \\ \vdots & \vdots \\ 0 & a_D \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

We set $\boldsymbol{b} := (\underbrace{0, \dots, 0}_{L_0 - 1 \text{ times}}, t)$. Then, $\boldsymbol{w}$ and $\boldsymbol{b}$ satisfy the condition of the lemma. □

C.3. Transformation of a Linear CNN into a ReLU CNN

The following lemma shows that we can convert any linear CNN to a ReLU CNN that has approximately 4 times larger parameters. This type of lemma is also found in Petersen & Voigtlaender (2018b) (Lemma 2.3). The constructed network resembles to a CNN with CReLU activation (Shang et al., 2016).

Lemma 2. *Let $\boldsymbol{C} = (C^{(1)}, \dots, C^{(L)}) \in \mathbb{N}_+^L$ be channel sizes $\boldsymbol{K} = (K^{(1)}, \dots, K^{(L)}) \in \mathbb{N}_+^L$ be filter sizes. Let $w^{(l)} \in \mathbb{R}^{K^{(l)} \times C^l \times C^{(l)}}$ and $b^{(l)} \in \mathbb{R}^{(l)}$. Consider the linear convolution layers constructed from $\boldsymbol{w}$ and $\boldsymbol{b}$: $f_{\mathrm{id}} := \mathrm{Conv}^{\mathrm{id}}_{\boldsymbol{w},\boldsymbol{b}} : \mathbb{R}^D \to \mathbb{R}^{D \times C^{(L)}} \mathbb{N}_+^L$ where $\boldsymbol{w} = (w^{(l)})_l$ and $\boldsymbol{b} = (b^{(l)})_l$. Then, there exists a pair $\tilde{\boldsymbol{w}} = (\tilde{w}^{(l)})_{l \in [L]}, \tilde{\boldsymbol{b}} = (\tilde{b}^{(l)})_{l \in [L]}$ where $\tilde{w}^{(l)} \in \mathbb{R}^{K^{(l)} \times 2C^{(l)} \times 2C^{(l-1)}}$ and $\tilde{b}^{(l)} \in \mathbb{R}^{2C^{(l)}}$ such that*

1. $\max_{l\in[L]}\|\tilde{w}^{(l)}\|_\infty = \max_{l\in[L]}\|w^{(l)}\|_\infty$, $\max_{l\in[L]}\|\tilde{b}^{(l)}\|_\infty = \max_{l\in[L]}\|b^{(l)}\|_\infty$, *and*
2. $f_{\mathrm{ReLU}} := \mathrm{Conv}^{\mathrm{ReLU}}_{\tilde{\boldsymbol{w}},\tilde{\boldsymbol{b}}} : \mathbb{R}^D \to \mathbb{R}^{D\times 2C^{(L)}}$, *satisfies* $f_{\mathrm{ReLU}}(\cdot) = (f_{\mathrm{id}}(\cdot)_+, f_{\mathrm{id}}(\cdot)_-)$.

Proof. We define $\tilde{\boldsymbol{w}}$ and $\tilde{\boldsymbol{b}}$ as follows:

$$
\begin{aligned}
(\tilde{w}^{(1)})_{k,:,:} &= \begin{bmatrix} (w^{(1)})_{k,:,:} \\ -(w^{(1)})_{k,:,:} \end{bmatrix} \text{ for } k = 1,\dots,K^{(1)}, \\
(\tilde{w}^{(l)})_{k,:,:} &= \begin{bmatrix} (w^{(l)})_{k,:,:} & -(w^{(l)})_{k,:,:} \\ -(w^{(l)})_{k,:,:} & (w^{(l)})_{k,:,:} \end{bmatrix} \text{ for } k = 1,\cdots K^{(l)}, \\
\tilde{b}^{(l)} &= \begin{bmatrix} b^{(l)} \\ -b^{(l)} \end{bmatrix}
\end{aligned}
$$

By definition, a pair $(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{b}})$ satisfies the conditions (1) and (2). For any $x\in\mathbb{R}^D$, we set $y^{(l)} := \mathrm{Conv}^{\mathrm{id}}_{\boldsymbol{w}[1:l],\boldsymbol{b}[1:l]}(x) \in \mathbb{R}^{C^{(l)}\times D}$. We will prove

$$
\mathrm{Conv}^{\mathrm{ReLU}}_{\tilde{\boldsymbol{w}}[1:l],\tilde{\boldsymbol{b}}[1:l]}(x) = \begin{bmatrix} y^{(l)}_+ & y^{(l)}_- \end{bmatrix}^\top \tag{1}
$$

for $l = 1,\dots,L$ by induction. Note that we obtain $f_{\mathrm{ReLU}}(\cdot) = (f_{\mathrm{id}+}(\cdot), f_{\mathrm{id}-}(\cdot))$ by setting $l = L$. For $l = 1$, by definition of $\tilde{w}^{(1)}$ we have,

$$
(\tilde{w}^{(1)})_{\alpha,:,:}x^{\beta,:} = \begin{bmatrix} (w^{(1)})_{\alpha,:,:}x^{\beta,:} \\ -(w^{(1)})_{\alpha,:,:}x^{\beta,:} \end{bmatrix}
$$

for any $\alpha,\beta\in[D]$. Summing them up and using the definition of $\tilde{b}^{(1)}$ yield

$$
[L^{\tilde{w}^{(1)}}(x) - \mathbf{1}_D\otimes\tilde{b}^{(1)}]^\top = \begin{bmatrix} L^{w^{(1)}}(x) - \mathbf{1}_D\otimes b^{(1)} \\ -\left(L^{w^{(1)}}(x) - \mathbf{1}_D\otimes b^{(1)}\right) \end{bmatrix}^\top
$$

Suppose (1) holds up to l ($l < L$), by the definition of $\tilde{w}^{(l+1)}$,

$$
\begin{aligned}
(\tilde{w}^{(l+1)})_{\alpha,:,:}\begin{bmatrix} (y^{(l)}_+)^{\beta,:} \\ (y^{(l)}_-)^{\beta,:} \end{bmatrix} &= \begin{bmatrix} (w^{(l+1)})_{\alpha,:,:} & -(w^{(l+1)})_{\alpha,:,:} \\ -(w^{(l+1)})_{\alpha,:,:} & (w^{(l+1)})_{\alpha,:,:} \end{bmatrix}\begin{bmatrix} (y^{(l)}_+)^{\beta,:} \\ (y^{(l)}_-)^{\beta,:} \end{bmatrix} \\
&= \begin{bmatrix} (w^{(l+1)})_{\alpha,:,:}\left((y^{(l)}_+)^{\beta,:} - (y^{(l)}_-)^{\beta,:}\right) \\ -(w^{(l+1)})_{\alpha,:,:}\left((y^{(l)}_+)^{\beta,:} - (y^{(l)}_-)^{\beta,:}\right) \end{bmatrix} \\
&= \begin{bmatrix} (w^{(l+1)})_{\alpha,:,:}(y^{(l)})^{\beta,:} \\ -(w^{(l+1)})_{\alpha,:,:}(y^{(l)})^{\beta,:} \end{bmatrix}
\end{aligned}
$$

for any $\alpha,\beta\in[D]$. Again, by taking the summation and using the definition of $\tilde{b}^{(l+1)}$, we get

$$
[L^{\tilde{w}^{(l+1)}}([y^{(l)}_+, y^{(l)}_-]) - \mathbf{1}_D\otimes\tilde{b}^{(1)}]^\top = \begin{bmatrix} L^{w^{(l+1)}}(y^{(l)}) - \mathbf{1}_D\otimes b^{(l+1)} \\ -\left(L^{w^{(l+1)}}(y^{(l)}) - \mathbf{1}_D\otimes b^{(l+1)}\right) \end{bmatrix}^\top.
$$

By applying ReLU, we get

$$
\mathrm{Conv}^{\mathrm{ReLU}}_{\tilde{w}^{(l+1)},\tilde{b}^{(l+1)}}\left([y^{(l)}_+, y^{(l)}_-]\right) = \mathrm{ReLU}\left([y^{(l+1)}, -y^{(l+1)}]\right). \tag{2}
$$

By using the induction hypothesis, we get

$$
\begin{aligned}
\mathrm{Conv}^{\mathrm{ReLU}}_{\tilde{\boldsymbol{w}}[1:(l+1)],\tilde{\boldsymbol{b}}[1:(l+1)]}(x) &= \mathrm{Conv}^{\mathrm{ReLU}}_{\tilde{w}^{(l+1)},\tilde{b}^{(l+1)}}\left([y^{(l)}_+, y^{(l)}_-]\right) \\
&= \mathrm{ReLU}\left([y^{(l+1)}, -y^{(l+1)}]\right) \\
&= [y^{(l+1)}_+, -y^{(l+1)}_-]
\end{aligned}
$$

Therefore, the claim holds for $l+1$. By induction, the claim holds for L, which is what we want to prove. □

C.4. Concatenation of CNNs

We can concatenate two CNNs with the same depths and filter sizes in parallel. Although it is almost trivial, we state it formally as a proposition. In the following proposition, $C^{(0)}$ and $C'^{(0)}$ is not necessarily 1.

Proposition 1. *Let $\boldsymbol{C} = (C^{(l)})_{l\in[L]}$, $\boldsymbol{C}' = (C'^{(l)})_{l\in[L]}$, and $\boldsymbol{K} = (K^{(l)})_{l\in[L]} \in \mathbb{N}_+^L$. Let $w^{(l)} \in \mathbb{R}^{K^{(l)}\times C^{(l)}\times C^{(l-1)}}$, $b \in \mathbb{R}^{C^{(l)}}$ and denote $\boldsymbol{w} = (w^{(l)})_l$ and $\boldsymbol{b} = (b^{(l)})_l$. We define $\boldsymbol{w}'$ and $\boldsymbol{b}'$ in the same way, with the exception that $C^{(l)}$ is replaced with $C'^{(l)}$. We define $\tilde{\boldsymbol{w}} = (\tilde{w}^{(1)}, \dots, \tilde{w}^{(L)})$ and $\tilde{\boldsymbol{b}} = (\tilde{b}^{(1)}, \dots, \tilde{b}^{(L)})$ by*

$$(\tilde{w}^{(l)})_{k,:,:} := \begin{bmatrix} w^{(l)} & 0 \\ 0 & w'^{(l)} \end{bmatrix} \in \mathbb{R}^{(C^{(l)}+C'^{(l)})\times(C^{(l-1)}+C'^{(l-1)})}$$
$$\tilde{b}^{(l)} := \begin{bmatrix} b^{(l)} \\ b'^{(l)} \end{bmatrix} \in \mathbb{R}^{(C^{(l)}+C'^{(l)})}$$

for $l \in [L]$ and $k \in [K^{(l)}]$. Then, we have,

$$\mathrm{Conv}^{\sigma}_{\tilde{\boldsymbol{w}},\tilde{\boldsymbol{b}}}(\begin{bmatrix} x & x' \end{bmatrix}) = \begin{bmatrix} \mathrm{Conv}^{\sigma}_{\boldsymbol{w},\boldsymbol{b}}(x) & \mathrm{Conv}^{\sigma}_{\boldsymbol{w}',\boldsymbol{b}'}(x') \end{bmatrix}$$

for any $x, x' \in \mathbb{R}^{D\times C^{(0)}}$ and any $\sigma : \mathbb{R} \to \mathbb{R}$. □

Note that by the definition of $\|\cdot\|_0$ and $\|\cdot\|_\infty$, we have

$$\max_{l\in[L]} \|\tilde{w}^{(l)}\|_\infty = \max_{l\in[L]} \|w^{(l)}\|_\infty \vee \|w'^{(l)}\|_\infty, \quad \text{and}$$
$$\max_{l\in[L]} \|\tilde{b}^{(l)}\|_\infty = \max_{l\in[L]} \|b^{(l)}\|_\infty \vee \|b'^{(l)}\|_\infty.$$

C.5. Proof of Theorem 5

By the definition of $\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$, there exists a 4-tuple $\boldsymbol{\theta} = ((W_m^{(l)})_{m,l}, (b_m^{(l)})_{m,l}, (w_m)_m, b)$ compatible with $(D_m^{(l)})_{m,l}$ ($m \in [M]$ and $l \in [L_m]$) such that

$$\max_{m\in[M],l\in[L_m]} (\|W_m^{(l)}\|_\infty \vee \|b_m^{(l)}\|_\infty) \le B^{(\mathrm{bs})},$$
$$\max_{m\in[M]} \|w_m\|_\infty \vee |b| \le B^{(\mathrm{fin})},$$

and $f^{(\mathrm{FNN})} = \mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}}$. We will construct the desired CNN consisting of M residual blocks, whose m-th residual block is made from the ingredients of the corresponding m-th block in $f^{(\mathrm{FNN})}$ (specifically, $\boldsymbol{W}_m := (W_m^{(l)})_{l\in[L_m]}$, $\boldsymbol{b}_m := (b_m^{(l)})_{l\in[L_m]}$, and w_m).

[Padding Block]: We prepare the padding operation P that multiply the channel size by 3 (i.e., we set $C^{(0)} = 3$).

[$\boldsymbol{m = 1, \dots, M}$ Blocks]: For fixed $m \in [M]$, we first create a CNN realizing $\mathrm{FC}^{\mathrm{ReLU}}_{\boldsymbol{W}_m,\boldsymbol{b}_m}$. We treat the first layer (i.e. $l = 1$) of $\mathrm{FC}^{\mathrm{ReLU}}_{\boldsymbol{W}_m,\boldsymbol{b}_m}$ as concatenation of $D_m^{(1)}$ hinge functions $\mathbb{R}^D \ni x \mapsto f_d(x) := ((W_m^{(1)})_d x - b_m^{(1)})_+$ for $d \in [D_m^{(1)}]$. Here, $(W_m^{(1)})_d \in \mathbb{R}^{1\times D}$ is the d-th row of the matrix $W_m^{(1)} \in \mathbb{R}^{D_m^{(1)}\times D}$. We apply Lemma 1 and Lemma 2 and obtain ReLU CNNs realizing the hinge functions. By combining them in parallel using Proposition 1, we have a learnable parameter $\boldsymbol{\theta}_m^{(1)}$ such that the ReLU CNN $\mathrm{Conv}^{\mathrm{ReLU}}_{\boldsymbol{\theta}_m^{(1)}} : \mathbb{R}^{D\times 2} \to \mathbb{R}^{D\times 2D_m^{(1)}}$ constructed from $\boldsymbol{\theta}_m^{(1)}$ satisfies

$$\mathrm{Conv}^{\mathrm{ReLU}}_{\boldsymbol{\theta}_m^{(1)}}(\begin{bmatrix} x & x' \end{bmatrix}^\top)_1 = \begin{bmatrix} f_1(x) & * & \cdots & f_{D_m^{(1)}}(x) & * \end{bmatrix}^\top.$$

Since we double the channel size in the $m = 0$ part, the identity mapping has 2 channels. Therefore, we made $\mathrm{Conv}^{\mathrm{ReLU}}_{\boldsymbol{\theta}_m^{(1)}}$ so that it has 2 input channels and neglects the input signals coming from the second one. This is possible by adding filters consisting of zeros appropriately.

Next, for l-th layer ($l = 2, \dots, L_m$), we prepare size-1 filters $w_m^{(2)} \in \mathbb{R}_m^{1\times D_m^{(2)}\times 2D^{(1)}}$ for $l = 2$ and $w_m^{(l)} \in \mathbb{R}^{1\times D_m^{(l)}\times 2D_m^{(l-1)}}$ for $l = 3, \dots, D_m^{(L_m)}$ defined by

$$(w_m^{(l)})_{1,:,:} := \begin{cases} W_m^{(2)} \otimes \begin{bmatrix} 1 & 0 \end{bmatrix} & \text{if } l = 2 \\ W_m^{(l)} & \text{if } l = 3, \dots, D_m^{(L_m)}, \end{cases}$$

where $\otimes$ is the Kronecker product of matrices. Intuitively, the $l = 2$ layer will pick all odd indices of the output of $\mathrm{Conv}_{\boldsymbol{\theta}_m^{(1)}}^{\mathrm{ReLU}}$ and apply the fully-connected layer. We construct CNNs from $\theta_m^{(l)} := (w_m^{(l)}, b_m^{(l)})$ ($l \geq 2$) and concatenate them along with $\mathrm{Conv}_{\boldsymbol{\theta}_m^{(1)}}^{\mathrm{ReLU}}$:

$$\mathrm{Conv}_m := \mathrm{Conv}_{\theta_m^{(L_m)}}^{\mathrm{ReLU}} \circ \cdots \circ \mathrm{Conv}_{\theta_m^{(2)}}^{\mathrm{ReLU}} \circ \mathrm{Conv}_{\boldsymbol{\theta}_m^{(1)}}^{\mathrm{ReLU}}.$$

Note that $\mathrm{Conv}_{\theta_m^{(l)}}^{\mathrm{ReLU}}$ ($l \geq 2$) just rearranges parameters of $\mathrm{FC}_{\boldsymbol{W}_m, \boldsymbol{b}_m}^{\mathrm{ReLU}}$. The output dimension of Conv_m is either $\mathbb{R}^{D\times 2D_m^{(L_m)}}$ (if $L_m = 1$) or $\mathbb{R}^{D\times D_m^{(L_m)}}$ (if $L_m \geq 2$)., We denote the output channel size (either $2D_m^{(L_m)}$ or $D_m^{(L_m)}$) by $D_m^{(\mathrm{out})}$. By the inductive calculation, we have

$$\mathrm{Conv}_m(x)_1 = \begin{cases} \mathrm{FC}_{\boldsymbol{W}_m, \boldsymbol{b}_m}^{\mathrm{ReLU}}(x) \otimes \begin{bmatrix} 1 & 0 \end{bmatrix} & \text{if } L_m = 1 \\ \mathrm{FC}_{\boldsymbol{W}_m, \boldsymbol{b}_m}^{\mathrm{ReLU}}(x) & \text{if } L_m \geq 2 \end{cases}.$$

By definition, Conv_m has $L_0 + L_m - 1$ layers and at most $4D_m^{(1)} \vee \max_{l=2,\dots L_m} D_m^{(l)} \leq 4 \max_{l\in[L_m]} D_m^{(l)}$ channels. The ∞-norm of its parameters does not exceed that of parameters in $\mathrm{FC}_{\boldsymbol{W}_m, \boldsymbol{b}_m}^{\mathrm{ReLU}}$.

Next, we consider the filter $\tilde{w}_m \in \mathbb{R}^{1\times 3\times D_m^{(\mathrm{out})}}$ defined by

$$(\tilde{w}_m)_{1,:,:} = \frac{B^{(\mathrm{bs})}}{B^{(\mathrm{fin})}} \begin{cases} \begin{bmatrix} 0 & \cdots & 0 \\ w_m \otimes \begin{bmatrix} 0 & 1 \end{bmatrix} \\ -w_m \otimes \begin{bmatrix} 0 & 1 \end{bmatrix} \end{bmatrix} & \text{if } L_m = 1 \\ \begin{bmatrix} 0 & \cdots & 0 \\ w_m & & \\ -w_m & & \end{bmatrix} & \text{if } L_m \geq 2 \end{cases},$$

Then, $\mathrm{Conv}'_m := \mathrm{Conv}_{\tilde{w}_m, 0}^{\mathrm{ReLU}}$ adds the output of m-th residual block, weighted by w_m, to the second channel in the identity connections, while keeping the first channel intact. Note that the final layer of each residual block does not have the ReLU activation. By definition, Conv'_m has $D_m^{(L_m)}$ parameters.

Given Conv_m and Conv'_m for each $m \in [M]$, we construct a CNN realizing $\mathrm{FNN}_{\boldsymbol{\theta}}^{\mathrm{ReLU}}$. Let $f^{(\mathrm{conv})} : \mathbb{R}^D \to \mathbb{R}^{D\times 3}$ be the sequential interleaving concatenation of Conv_m and Conv'_m, that is,

$$f^{(\mathrm{conv})} := (\mathrm{Conv}'_M \circ \mathrm{Conv}_M + I) \circ \cdots \quad \circ (\mathrm{Conv}'_1 \circ \mathrm{Conv}_1 + I) \circ P.$$

Then, we have

$$f_{1,:}^{(\mathrm{conv})} = \begin{bmatrix} 0 & z_1 & z_2 \end{bmatrix} \in \mathbb{R}^3$$

where $z_1 = \frac{B^{(\mathrm{bs})}}{B^{(\mathrm{fin})}} \sum_{m=1}^M \left(w_m^\top \mathrm{FC}_{\boldsymbol{W}_m, \boldsymbol{b}_m}^{\mathrm{ReLU}}\right)_+$ and $z_2 = \frac{B^{(\mathrm{bs})}}{B^{(\mathrm{fin})}} \sum_{m=1}^M \left(w_m^\top \mathrm{FC}_{\boldsymbol{W}_m, \boldsymbol{b}_m}^{\mathrm{ReLU}}\right)_-$.

[Final Fully-connected Layer] Finally, we set

$$w := \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \frac{B^{(\mathrm{fin})}}{B^{(\mathrm{bs})}} & 0 & \cdots & 0 \\ -\frac{B^{(\mathrm{fin})}}{B^{(\mathrm{bs})}} & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{D\times 3}$$

and put $\mathrm{FC}^{\mathrm{id}}_{w,b}$ on top of $f^{(\mathrm{conv})}$ to pick the first coordinate of $f^{(\mathrm{conv})}$ and subtract the bias term. By definition, $f^{(\mathrm{CNN})} := \mathrm{FC}^{\mathrm{id}}_{w,b} \circ f^{(\mathrm{conv})}$ satisfies $f^{(\mathrm{CNN})} = f^{(\mathrm{FNN})}$.

[Property Check] We will check $f^{(\mathrm{FNN})}$ satisfies the desired properties. **(Property 1)**: Since Conv_m and Conv'_m has $L_0 + L_m - 1$ and 1 layers, respectively, the $m(\geq 1)$-th residual block of $f^{(\mathrm{CNN})}$ has $L'_m = L_0 + L_m$ layers. In particular, if L_m's are same, we can choose L'_m to be the same value $L_0 + L_m$. **(Property 2)**: Conv_m has at most $4 \max_{l \in [L_m]} D^{(l)}_m$ channels and Conv'_m has at most 2 channels, respectively. Therefore, the channel size of the m-th block is at most $4 \max_{l \in [L_m]} D^{(l)}_m$. **(Property 3)**: Since each filter of Conv_m and Conv'_m is at most K, the filter size of CNN is also at most K. **(Properties on $B^{(\mathrm{conv})}$ and $B^{(\mathrm{fin})}$)**: Parameters of $f^{(\mathrm{conv})}$ are either 0, or parameters of $\mathrm{FC}^{\mathrm{ReLU}}_{\boldsymbol{W}_m, \boldsymbol{W}_m}$, whose absolute value is bounded by $B^{(\mathrm{bs})}$ or $\frac{B^{(\mathrm{bs})}}{B^{(\mathrm{fin})}} w_m$. Since we have $\|w_m\|_\infty \leq B^{(\mathrm{fin})}$, the ∞-norm of parameters in $f^{(\mathrm{CNN})}$ is bounded by $B^{(\mathrm{bs})}$. The parameters of the final fully-connected layer $\mathrm{FC}_{w,b}$ is either $\frac{B^{(\mathrm{fin})}}{B^{(\mathrm{bs})}}$, 0, or b, therefore their norm is bounded by $\frac{B^{(\mathrm{fin})}}{B^{(\mathrm{bs})}} \vee B^{(\mathrm{fin})}$. □

As discussed in the beginning of this section, Theorem 1 is the special case of Theorem 5.

Remark 2. *Another way to construct a CNN which is identical (as a function) to a given FNN is as follows. First, we use a "rotation" convolution with D filters, each of which has a size D, to serialize all input signals to channels of a single input dimension. Then, apply size-1 convolution layers, whose l-th layer consisting of appropriately arranged weight parameters of the l-th layer of the FNN. This is essentially what Petersen & Voigtlaender (2018a) did to prove the existence of a CNN equivalent to a given FNN. To restrict the size of filters to K, we should further replace the the first convolution layer with $O(D/K)$ convolution layers with size-K filters. We can show essentially same statement using this construction method.*

D. Proof of Theorem 2

Same as Theorem 1, we restate Theorem 2 in more general form. We denote $\mathcal{F}^{(\mathrm{CNN})} := \mathcal{F}^{(\mathrm{CNN})}_{\boldsymbol{C}, \boldsymbol{K}, B^{(\mathrm{conv})}, B^{(\mathrm{fc})}}$ and $\mathcal{F}^{(\mathrm{FNN})} := \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D}, B^{(\mathrm{bs})}, B^{(\mathrm{fin})}}$ in shorthand.

Theorem 6. *Let $f^\circ : \mathbb{R}^D \to \mathbb{R}$ be a measurable function and $B^{(\mathrm{bs})}, B^{(\mathrm{fin})} > 0$. Let M, K, L_0, L_m, and $\boldsymbol{D}$ as in Theorem 5. Suppose $L'_m, \boldsymbol{C}, \boldsymbol{K}, B^{(\mathrm{conv})}$ and $B^{(\mathrm{fc})}$ satisfy $\mathcal{F}^{(\mathrm{FNN})} \subset \mathcal{F}^{(\mathrm{CNN})}$ for $B^{(\mathrm{bs})}$ and $B^{(\mathrm{fin})}$ (their existence is ensured for any $B^{(\mathrm{bs})}$ and $B^{(\mathrm{fin})}$ by Theorem 5). Suppose that the covering nubmer of $\mathcal{F}^{(\mathrm{CNN})}$ is larger than 3. Then, the clipped ERM estimator $\hat{f}$ in $\mathcal{F} := \{\mathrm{clip}[f] \mid f \in \mathcal{F}^{(\mathrm{CNN})}\}$ satisfies*

$$\mathbb{E}_{\mathcal{D}} \|\hat{f} - f^\circ\|^2_{\mathcal{L}^2(\mathcal{P}_X)} \leq C \left(\inf_f \|f - f^\circ\|^2_\infty + \frac{\tilde{F}^2}{N} \Lambda_2 \log(2\Lambda_1 B N) \right). \tag{3}$$

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

$$\Lambda_1 := (2M+3) C^{(0)} D (1 \vee B^{(\mathrm{fc})})(1 \vee B^{(\mathrm{conv})}) \varrho \varrho^+$$
$$\Lambda_2 := \sum_{m=1}^{M} \sum_{l=1}^{L'_m} \left(C^{(l-1)}_m C^{(l)}_m K^{(l)}_m + C^{(l)}_m \right) + C^{(0)} D + 1,$$

where $\varrho = \prod_{m=1}^M (1 + \rho_m)$, $\varrho^+ = 1 + \sum_{m=1}^M L'_m \rho^+_m$, $\rho_m := \prod_{l=1}^{L'_m} C^{(l-1)}_m K^{(l)}_m B^{(\mathrm{conv})}$ and $\rho^+_m := \prod_{l=1}^{L'_m} (1 \vee C^{(l-1)}_m K^{(l)}_m B^{(\mathrm{conv})})$.

Again, Theorem 2 is a special case of Theorem 6 where width, depth, channel sizes and filter sizes are same among blocks. Note that the definitions of Λ_1, Λ_2, ρ, ρ^+, ϱ, and ϱ^+ in Theorem 2 and Theorem 6 are consistent by this specialization.

D.1. Proof Overview

We relate the approximation error of Theorem 2 with the estimation error using the covering number of the hypothesis class $\mathcal{F}^{(\mathrm{CNN})}$. Although there are several theorems of this type, we employ the one in Schmidt-Hieber (2017) due to its convenient form (Lemma 5). We can prove that the logarithm of the covering number is upper bounded by $\Lambda_2 \log((B^{(\mathrm{conv})} \vee B^{(\mathrm{fc})}) \Lambda_1 / \varepsilon)$

(Lemma 4) using the similar techniques to the one in Schmidt-Hieber (2017). Theorem 2 is the immediate consequence of these two lemmas.

To prove Corollary 1, we set $M = O(N^\alpha)$ for some $\alpha > 0$. Then, under the assumption of the corolarry, we have $\|f^\circ - \hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_x)} = \tilde{O}\left(\max\left(N^{-2\alpha\gamma_1}, N^{\alpha\gamma_2 - 1}\right)\right)$ from Theorem 2. The order of the right hand side with respect to N is minimized when $\alpha = \frac{1}{2\gamma_1+\gamma_2}$. By substituting α, we can prove Corollary 1.

D.2. Covering Number of CNNs

The goal of this section is to prove Lemma 4, stated in Section D.2.5, that evaluates the covering number of the set of functions realized by CNNs.

D.2.1. Bounds for convolutional layers

We assume $w, w' \in \mathbb{R}^{K\times J\times I}$, $b, b' \in \mathbb{R}$, and $x \in \mathbb{R}^{D\times I}$ unless specified. We have in mind that the activation function σ is either the ReLU function or the identity function id. But the following proposition holds for any 1-Lipschitz function such that $\sigma(0) = 0$. Remember that we can treat L^w as a linear operator from $\mathbb{R}^{D\times I}$ to $\mathbb{R}^{D\times J}$. We endow $\mathbb{R}^{D\times I}$ and $\mathbb{R}^{D\times J}$ with the sup norm and denote the operator norm L^w by $\|L^w\|_{\mathrm{op}}$.

Proposition 2. *It holds that* $\|L^w\|_{\mathrm{op}} \le IK\|w\|_\infty$.

Proof. Write $w = (w_{kji})_{k\in[K],j\in[J],i\in[I]}$, $L^w = ((L^w)^{\beta,j}_{\alpha,i})_{\alpha,\beta\in[D],j\in[J],i\in[I]}$. For any $x = (x_{\alpha,i})_{\alpha\in[D],i\in[I]} \in \mathbb{R}^{D\times I}$, the sup norm of $y := (y_{\beta j})_{\beta\in[D]j\in[J]} = L^w(x)$ is evaluated as follows:

$$\begin{aligned}
\|y\|_\infty = \max_{\beta,j}|y_{\beta,j}| &\le \max_{\beta,j}\sum_{\alpha,i}|(L^w)^{\beta,j}_{\alpha,i}||x_{\alpha,i}| \\
&\le \max_{\beta,j}\sum_{\alpha,i}|(L^w)^{\beta,j}_{\alpha,i}|\|x\|_\infty \\
&= \max_{\beta,j}\sum_{\alpha,i}|w_{(\alpha-\beta+1),j,i}|\|x\|_\infty \\
&\le IK\|w\|_\infty\|x\|_\infty
\end{aligned}$$

□

Proposition 3. *It holds that* $\|\mathrm{Conv}^\sigma_{w,b}(x)\|_\infty \le \|L^w\|_{\mathrm{op}}\|x\|_\infty + |b|$.

Proof.

$$\begin{aligned}
\|\mathrm{Conv}^\sigma_{w,b}(x)\|_\infty &\le \|\sigma(L^w(x) - \mathbf{1}_D \otimes b)\|_\infty \\
&\le \|L^w(x) - \mathbf{1}_D \otimes b\|_\infty \\
&\le \|L^w(x)\|_\infty + \|\mathbf{1}_D \otimes b\|_\infty \\
&\le \|L^w\|_{\mathrm{op}}\|x\|_\infty + |b|.
\end{aligned}$$

□

Proposition 4. *The Lipschitz constant of* $\mathrm{Conv}^\sigma_{w,b}$ *is bounded by* $\|L^w\|_{\mathrm{op}}$.

Proof. For any $x, x' \in \mathbb{R}^{D\times I}$,

$$\begin{aligned}
\|\mathrm{Conv}^\sigma_{w,b}(x) - \mathrm{Conv}^\sigma_{w,b}(x')\|_\infty &= \|\sigma\left(L^w(x) - \mathbf{1}_D \otimes b\right) - \sigma\left(L^w(x') - \mathbf{1}_D \otimes b\right)\|_\infty \\
&\le \|\left(L^w(x) - \mathbf{1}_D \otimes b\right) - \left(L^w(x') - \mathbf{1}_D \otimes b\right)\|_\infty \\
&\le \|L^w(x - x')\|_\infty \\
&\le \|L^w\|_{\mathrm{op}}\|x - x'\|_\infty.
\end{aligned}$$

Note that the first inequality holds because the ReLU function is 1-Lipschitz. □

Proposition 5. *It holds that* $\|\mathrm{Conv}^{\sigma}_{w,b}(x) - \mathrm{Conv}^{\sigma}_{w',b'}(x)\| \le \|L^{w-w'}\|_{\mathrm{op}}\|x\|_\infty + |b-b'|$.

Proof.

$$\begin{aligned}\|\mathrm{Conv}^{\sigma}_{w,b}(x) - \mathrm{Conv}^{\sigma}_{w',b'}(x)\|_\infty &= \|\sigma(L^w(x) - \mathbf{1}_D \otimes b) - \sigma(L^{w'}(x) - \mathbf{1}_D \otimes b')\|_\infty \\ &\le \|(L^w(x) - \mathbf{1}_D \otimes b) - (L^{w'}(x) - \mathbf{1}_D \otimes b')\| \\ &= \|L^w(x) - L^{w'}(x)\| + \|\mathbf{1}_D \otimes (b-b')\|_\infty \\ &\le \|L^{w-w'}\|_{\mathrm{op}}\|x\|_\infty + |b-b'|\end{aligned}$$

□

D.2.2. Bounds for fully-connected layers

In the following propositions in this subsection, we assume $W, W' \in \mathbb{R}^{DC\times C'}$, $b, b' \in \mathbb{R}^{C'}$, and $x \in \mathbb{R}^{D\times C}$. Again, these propositions hold for any 1-Lipschitz function $\sigma : \mathbb{R} \to \mathbb{R}$ such that $\sigma(0) = 0$. But $\sigma = \mathrm{ReLU}$ or id is enough for us.

Proposition 6. *It holds that* $\|\mathrm{FC}^{\sigma}_{W,b}(x)\|_\infty \le \|W\|_0\|W\|_\infty\|x\|_\infty + \|b\|_\infty$.

Proof.

$$\|\mathrm{FC}^{\sigma}_{W,b}(x)\|_\infty \le \|W\mathrm{vec}(x) - b\|_\infty \le \|W\mathrm{vec}(x)\|_\infty + \|b\|_\infty \le \max_j \sum_{\alpha,i} \left|W_{\alpha,i,j}x^{\alpha,i}\right| + \|b\|_\infty.$$

The number of non-zero summand in the summation is at most $\|W\|_0$ and each summand is bounded by $\|W\|_\infty\|x\|_\infty$ Therefore, we have $\|\mathrm{FC}^{\sigma}_{W,b}(x)\|_\infty \le \|W\|_0\|W\|_\infty\|x\|_\infty + \|b\|_\infty$. □

Proposition 7. *The Lipschitz constant of* $\mathrm{FC}^{\sigma}_{W,b}$ *is bounded by* $\|W\|_0\|W\|_\infty$.

Proof. For any $x, x' \in \mathbb{R}^{D\times C}$,

$$\begin{aligned}\|\mathrm{FC}^{\sigma}_{W,b}(x) - \mathrm{FC}^{\sigma}_{W,b}(x')\|_\infty &\le \|(W\mathrm{vec}(x) - b) - (W\mathrm{vec}(x') - b)\|_\infty \\ &\le \|W(\mathrm{vec}(x) - \mathrm{vec}(x'))\|\infty \\ &\le \|W\|_0\|W\|_\infty\|\mathrm{vec}(x) - \mathrm{vec}(x')\|_\infty.\end{aligned}$$

□

Proposition 8. *It holds that* $\|\mathrm{FC}^{\sigma}_{W,b}(x) - \mathrm{FC}^{\sigma}_{W',b'}(x)\|_\infty \le (\|W\|_0 + \|W'\|_0)\|W - W'\|_\infty\|x\|_\infty + \|b-b'\|_\infty$.

Proof.

$$\begin{aligned}\|\mathrm{FC}^{\sigma}_{W,b}(x) - \mathrm{FC}^{\sigma}_{W',b'}(x)\|_\infty &\le \|(W\mathrm{vec}(x) - b) - (W'\mathrm{vec}(x) - b')\|_\infty \\ &= \|((W-W')\mathrm{vec}(x) - (b-b')\|_\infty \\ &\le \|(W-W')\mathrm{vec}(x)| + \|b-b'\|_\infty \\ &\le \|W-W'\|_0\|W-W'\|_\infty\|x\|_\infty + \|b-b'\|_\infty \\ &\le (\|W\|_0 + \|W'\|_0)\|W-W'\|_\infty\|x\|_\infty + \|b-b'\|_\infty\end{aligned}$$

□

D.2.3. BOUNDS FOR RESIDUAL BLOCKS

In this section, we denote the architecture of CNNs by $\boldsymbol{C} = (C^{(l)})_{l\in[L]} \in \mathbb{N}_+^L$ and $\boldsymbol{K} = (K^{(l)})_{l\in[L]} \in \mathbb{N}_+^L$ and the norm constraint on the convolution part by $B^{(\mathrm{conv})}$ ($C^{(0)}$ need not equal to 1 in this section). Let $w^{(l)}, w'^{(l)} \in \mathbb{R}^{K^{(l)}\times C^{(l)}\times C^{(l-1)}}$ and $b^{(l)}, b'^{(l)} \in \mathbb{R}$. We denote $\boldsymbol{w} := (w^{(l)})_{l\in[L]}$, $\boldsymbol{b} := (b^{(l)})_{l\in[L]}$, $\boldsymbol{w}' := (w'^{(l)})_{l\in[L]}$, and $\boldsymbol{b} := (b^{(l)})_{l\in[L]}$.

For $1 \le l \le l' \le L$, we denote $\rho(l, l') := \prod_{i=l}^{l'}(C^{(i-1)}K^{(i)}B^{(\mathrm{conv})})$ and $\rho^+(l, l') := \prod_{i=l}^{l'} 1 \vee (C^{(i-1)}K^{(i)}B^{(\mathrm{conv})})$.

Proposition 9. *Let $l \in [L]$. We assume $\max_{l\in[L]} \|w^{(l)}\|_\infty \vee \|b^{(l)}\|_\infty \le B^{(\mathrm{conv})}$. Then, for any $x \in [-1, 1]^{D\times C^{(0)}}$, we have $\|\mathrm{Conv}^\sigma_{\boldsymbol{w}[1:l],\boldsymbol{b}[1:l]}(x)\|_\infty \le \rho(1, l)\|x\|_\infty + B^{(\mathrm{conv})} l\rho^+(1, l)$.*

Proof. We write in shorthand as $C_{[s:t]} := \mathrm{Conv}^\sigma_{\boldsymbol{w}[s:t],\boldsymbol{b}[s:t]}$. Using Proposition 3 recursively, we get

$$\begin{aligned}\|C_{[1:l]}(x)\|_\infty &\le \|L^{w^{(l)}}\|_{\mathrm{op}}\|C_{[1:l-1]}(x)\|_\infty + \|b^{(l)}\|_\infty \\ &\cdots \\ &\le \|x\|_\infty \prod_{i=1}^{l}\|L^{w^{(i)}}\|_{\mathrm{op}} + \sum_{i=2}^{l}\|b^{(i-1)}\|_\infty \prod_{j=i}^{l}\|L^{w^{(j)}}\|_{\mathrm{op}} + \|b^{(l)}\|_\infty.\end{aligned}$$

By Proposition 2 and assumptions $\|w^{(i)}\|_\infty \le B^{(\mathrm{conv})}$ and $\|b^{(i)}\|_\infty \le B^{(\mathrm{conv})}$, it is further bounded by

$$\begin{aligned}&\|x\|_\infty \prod_{i=1}^{l}(C^{(i-1)}K^{(i)}B^{(\mathrm{conv})}) + B^{(\mathrm{conv})}\sum_{i=2}^{l}\prod_{j=i}^{l}(C^{(j-1)}K^{(j)}B^{(\mathrm{conv})}) + B^{(\mathrm{conv})} \\ &\le \rho(1, l)\|x\|_\infty + B^{(\mathrm{conv})} l\rho^+(1, l)\end{aligned}$$

□

Proposition 10. *Let $\varepsilon > 0$, suppose $\max_{l\in[L]} \|w^{(l)} - w'^{(l)}\|_\infty \le \varepsilon$ and $\max_{l\in[L]} \|b^{(l)} - b'^{(l)}\|_\infty \le \varepsilon$, then $\|C_{[1:L]} - C'_{[1:L]}(x)\|_\infty \le (L\rho(1, L)\|x\|_\infty + (1 \vee B^{(\mathrm{conv})})L^2\rho^+(1, L))\varepsilon$ for any $x \in \mathbb{R}^{D\times C^{(0)}}$.*

Proof. For any $l \in [L]$, we have

$$\begin{aligned}&\left|C'_{[l+1:L]} \circ (C_l - C'_l) \circ C_{[1:l-1]}(x)\right| \\ &\le \|C'_{[l+1:L]} \circ (C_l - C'_l) \circ C_{[1:l-1]}(x)\|_\infty \\ &\le \rho(l + 1, L)\left\|(C_l - C'_l) \circ C_{[1:l-1]}(x)\right\|_\infty \quad \text{(by Proposition 2 and 4)} \\ &\le \rho(l + 1, L)\left(\rho(l, l)\|C_{[1:l-1]}\|_\infty\varepsilon + \varepsilon\right) \quad \text{(by Proposition 2 and 5)} \\ &\le \rho(l + 1, L)\left(\rho(l, l)(\rho(1, l - 1)\|x\|_\infty + B^{(\mathrm{conv})}(l - 1)\rho_+(1, l - 1)) + 1\right)\varepsilon \\ &\quad \text{(by Proposition 9)} \\ &= \left(\rho(1, L)\|x\|_\infty + (1 \vee B^{(\mathrm{conv})})L\rho_+(1, L)\right)\varepsilon \qquad (4)\end{aligned}$$

Therefore,

$$\begin{aligned}\|C_{[1:L]}(x) - C'_{[1:L]}(x)\|_\infty &\le \sum_{l=1}^{L}\|C_{[l+1:L]} \circ (C_l - C'_l) \circ C_{[1:l-1]}(x)\|_\infty \\ &\le (L\rho(1, L)\|x\|_\infty + (1 \vee B^{(\mathrm{conv})})L^2\rho^+(1, L))\varepsilon\end{aligned}$$

□

D.2.4. PUTTING THEM ALL

Let $M, L_m, C_m^{(l)}, K_m^{(l)} \in \mathbb{N}_+$, $\boldsymbol{C} := (C_m^{(l)})_{m,l}$, and $\boldsymbol{K} := (K_m^{(l)})_{m,l}$ for $m \in [M]$ and $l \in [L_m]$. Let $\boldsymbol{\theta} = ((w_m^{(l)})_{m,l}, (b_m^{(l)})_{m,l}, W, b)$ and $\boldsymbol{\theta}' = ((w'^{(l)}_m)_{m,l}, (b'^{(l)}_m)_{m,l}, W', b')$ be tuples compatible with $(\boldsymbol{C}, \boldsymbol{K})$ such that $\mathrm{CNN}_{\boldsymbol{\theta}}^{\mathrm{ReLU}}$, $\mathrm{CNN}_{\boldsymbol{\theta}'}^{\mathrm{ReLU}} \in \mathcal{F}^{(\mathrm{CNN})}_{\boldsymbol{C},\boldsymbol{K},B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$ for some $B^{(\mathrm{conv})}, B^{(\mathrm{fc})} > 0$. We denote the l-th convolution layer of the m-th block by $C_m^{(l)}$ and the m-th residual block of by C_m:

$$C_m^{(l)} := \begin{cases} \mathrm{Conv}^{\mathrm{id}}_{w_m^{(l)}} & (\text{if } l = L_m) \\ \mathrm{Conv}^{\mathrm{ReLU}}_{w_m^{(l)}} & (\text{otherwise}) \end{cases}$$
$$C_m := C_m^{(L_m)} \circ \cdots \circ C_m^{(1)}.$$

Also, we denote by $C_{[m:m']}$ the subnetwork of $\mathrm{Conv}_{\boldsymbol{\theta}}^{\mathrm{ReLU}}$ between the m-th and m'-th block. That is,

$$C_{[m:m']} := \begin{cases} (C_{m'} + I) \circ \cdots \circ (C_m + I) & (\text{if } m \geq 1) \\ (C_{m'} + I) \circ \cdots \circ (C_1 + I) \circ P & (\text{if } m = 0) \end{cases}$$

for $m, m' = 0, \ldots, M$. We define $C'^{(l)}_m$, C'_m and $C'_{[m:m']}$ similarly for $\boldsymbol{\theta}'$.

Proposition 11. *For $m \in [M]$ and $x \in [-1,1]^D$, we have $\|C_{[0:m]}(x)\|_\infty \leq (1 \vee B^{(\mathrm{conv})})\varrho_m \varrho_m^+$. Here, $\varrho_m = (\prod_{i=1}^m (1+\rho_i))$ and $\varrho_m^+ = \left(1 + \sum_{i=1}^m L_i \rho_i^+\right)$ (ρ_m and ρ_m^+ are constants defined in Theorem 6).*

Proof. By using Proposition 9 inductively, we have

$$\begin{aligned}
\|C_{[0:m]}(x)\|_\infty &\leq \|C_m(C_{[0:m-1]}(x)) + C_{[0:m-1]}(x)\|_\infty \\
&\leq \|(1+\rho_m)C_{[0:m-1]}(x) + B^{(\mathrm{conv})} L_m \rho_{+m})\|_\infty \\
&\leq (1+\rho_m)\|C_{[0:m-1]}(x)\|_\infty + B^{(\mathrm{conv})} L_m \rho_m^+ \\
&\cdots \\
&\leq \|P(x)\|_\infty \prod_{i=1}^m (1+\rho_i) + B^{(\mathrm{conv})} \sum_{i=1}^m L_i \rho_i^+ \prod_{j=i+1}^m (1+\rho_j) \\
&\leq \prod_{i=1}^m (1+\rho_i) + B^{(\mathrm{conv})} \sum_{i=1}^m L_i \rho_i^+ \prod_{j=i+1}^m (1+\rho_j) \\
&\leq (1 \vee B^{(\mathrm{conv})}) \varrho_m \varrho_m^+.
\end{aligned}$$

□

Lemma 3. *Let $\varepsilon > 0$. Suppose $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are within distance ε, that is, $\max_{m,l} \|w_m^{(l)} - w'^{(l)}_m\|_\infty \leq \varepsilon$, $\|b_m^{(l)} - b'^{(l)}_m\|_\infty \leq \varepsilon$, $\|W - W'\|_\infty \leq \varepsilon$, and $\|b - b'\|_\infty \leq \varepsilon$. Then, $\|\mathrm{CNN}_{\boldsymbol{\theta}}^{\mathrm{ReLU}} - \mathrm{CNN}_{\boldsymbol{\theta}'}^{\mathrm{ReLU}}\|_\infty \leq \Lambda_1 \varepsilon$ where Λ_1 is the constant defined in Theorem 6.*

Proof. For any $x \in [-1,1]^D$, we have

$$\begin{aligned}
\left|\mathrm{CNN}_{\boldsymbol{\theta}}^{\mathrm{ReLU}}(x) - \mathrm{CNN}_{\boldsymbol{\theta}'}^{\mathrm{ReLU}}(x)\right| &= \left|\mathrm{FC}^{\mathrm{id}}_{W,b} \circ C_{[0:M]}(x) - \mathrm{FC}^{\mathrm{id}}_{W',b'} \circ C'_{[0:M]}(x)\right| \\
&= \left|\left(\mathrm{FC}^{\mathrm{id}}_{W,b} - \mathrm{FC}^{\mathrm{id}}_{W',b'}\right) \circ C_{[0:M]}(x)\right| \\
&\quad + \sum_{m=1}^M \left|\mathrm{FC}^{\mathrm{id}}_{W',b'} \circ C_{[m+1:M]} \circ (C_m - C'_m) \circ C'_{[0:m-1]}(x)\right|. \qquad (5)
\end{aligned}$$

We will bound each term of (5). By Proposition 8 and Proposition 11,

$$\begin{aligned}\left|\left(\mathrm{FC}^{\mathrm{id}}_{W,b} - \mathrm{FC}^{\mathrm{id}}_{W',b'}\right) \circ C_{[0:M]}(x)\right| &\le (\|W\|_0 + \|W'\|_0)\|W - W'\|_\infty \|C_{[0:M]}(x)\|_\infty + \|b - b'\|_\infty \\ &\le 2C_0^{(L_0)} D \|C_{[0:M]}(x)\|_\infty \varepsilon + \varepsilon \\ &\le 2C_0^{(L_0)} D (1 \vee B^{(\mathrm{conv})}) \varrho_M \varrho_M^+ \varepsilon + \varepsilon \\ &\le 3C_0^{(L_0)} D (1 \vee B^{(\mathrm{conv})}) \varrho_M \varrho_M^+ \varepsilon. \end{aligned} \tag{6}$$

On the other hand, for $m \in [M]$,

$$\begin{aligned}&\left|\mathrm{FC}^{\mathrm{id}}_{W',b'} \circ C'_{[m+1:M]} \circ (C_m - C'_m) \circ C_{[0:m-1]}(x)\right| \\ &\le \|W'\|_0 \|W'\|_\infty \|C'_{[m+1:M]} \circ (C_m - C'_m) \circ C_{[1:m-1]}(x)\|_\infty \quad \text{(by Proposition 7)} \\ &\le C_0^{(L_0)} D B^{(\mathrm{fc})} \|C'_{[m+1:M]} \circ (C_m - C'_m) \circ C_{[0:m-1]}(x)\|_\infty \\ &\le C_0^{(L_0)} D B^{(\mathrm{fc})} \left(\prod_{i=m+1}^{M} \rho_i\right) \left\|(C_m - C'_m) \circ C_{[0:m-1]}(x)\right\|_\infty \quad \text{(by Proposition 2 and 4)} \\ &\le C_0^{(L_0)} D B^{(\mathrm{fc})} \left(\prod_{i=m+1}^{M} \rho_i\right) \left(\rho_m \|C_{[0:m-1]}\|_\infty \varepsilon + \varepsilon\right) \quad \text{(by Proposition 2 and 5)} \\ &\le C_0^{(L_0)} D B^{(\mathrm{fc})} \left(\prod_{i=m+1}^{M} \rho_i\right) \left(\rho_m (1 \vee B^{(\mathrm{conv})}) \varrho_{m-1} \varrho_{m-1}^+ + 1\right) \varepsilon \quad \text{(by Proposition 9)} \\ &\le 2C_0^{(L_0)} D B^{(\mathrm{fc})} (1 \vee B^{(\mathrm{conv})}) \varrho_M \varrho_M^+ \varepsilon \end{aligned} \tag{7}$$

By applying (6) and (7) to (5), we have

$$\begin{aligned}|\mathrm{CNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}}(x) - \mathrm{CNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}'}(x)| &\le 3C_0^{(L_0)} D (1 \vee B^{(\mathrm{conv})}) \varrho_M \varrho_M^+ \varepsilon \\ &\quad + 2MC_0^{(L_0)} D B^{(\mathrm{fc})} (1 \vee B^{(\mathrm{conv})}) \varrho_M \varrho_M^+ \varepsilon \\ &\le (2M+3) C_0^{(L_0)} D (1 \vee B^{(\mathrm{fc})}) (1 \vee B^{(\mathrm{conv})}) \varrho_M \varrho_M^+ \varepsilon \\ &= \Lambda_1 \varepsilon. \end{aligned}$$

□

D.2.5. Bounds for covering number of CNNs

For a metric space $(\mathcal{M}_0, d)$ and $\varepsilon > 0$, we denote the (external) covering number of $\mathcal{M} \subset \mathcal{M}_0$ by $\mathcal{N}(\varepsilon, \mathcal{M}, d)$: $\mathcal{N}(\varepsilon, \mathcal{M}, d) := \inf\{N \in \mathbb{N} \mid \exists f_1, \dots, f_N \in \mathcal{M_0} \text{ s.t. } \forall f \in \mathcal{M}, \exists n \in [N] \text{ s.t. } d(f, f_n) \le \varepsilon\}$.

Lemma 4. *Let $B := B^{(\mathrm{conv})} \vee B^{(\mathrm{fc})}$. For $\varepsilon > 0$, we have $\mathcal{N}(\varepsilon, \mathcal{F}^{(\mathrm{CNN})}, \|\cdot\|_\infty) \le \left(2B\Lambda_1 \varepsilon^{-1}\right)^{\Lambda_2}$.*

Proof. The idea of the proof is same as that of Lemma 5 of Schmidt-Hieber (2017). We divide the interval of each parameter range ($[-B^{(\mathrm{conv})}, B^{(\mathrm{conv})}]$ or $[-B^{(\mathrm{fc})}, B^{(\mathrm{fc})}]$) into bins with width $\Lambda_1^{-1}\varepsilon$ (i.e., $2B^{(\mathrm{conv})}\Lambda_1\varepsilon^{-1}$ or $2B^{(\mathrm{fc})}\Lambda_1\varepsilon^{-1}$ bins for each interval). If $f, f' \in \mathcal{F}^{(\mathrm{CNN})}$ can be realized by parameters such that every pair of corresponding parameters are in a same bin, then, $\|f - f'\|_\infty \le \varepsilon$ by Lemma 3. We make a subset $\mathcal{F}_0$ of $\mathcal{F}^{(\mathrm{CNN})}$ by picking up every combination of bins for Λ_2 parameters. Then, for each $f \in \mathcal{F}^{(\mathrm{CNN})}$, there exists $f_0 \in \mathcal{F}_0$ such that $\|f - f_0\|_\infty \le \varepsilon$. There are at most $2B\Lambda_1\varepsilon^{-1}$ choices of bins for each parameter. Therefore, the cardinality of $\mathcal{F}_0$ is at most $\left(2B\Lambda_1\varepsilon^{-1}\right)^{\Lambda_2}$. □

D.3. Proofs of Theorem 2 and Corollary 1

We use the lemma in Schmidt-Hieber (2017) to bound the estimation error of the clipped ERM estimator $\hat{f}$. Since our problem setting is slightly different from one in the paper, we restate the statement.

Lemma 5 (cf. Schmidt-Hieber (2017) Lemma 4)**.** *Let $\mathcal{F}$ be a family of measurable functions from $[-1,1]^D$ to $\mathbb{R}$. Let $\hat{f}$ be the clipped ERM estimator of the regression problem described in Section 3.1. Suppose the covering number of $\mathcal{F}$ satisfies $\mathcal{N}(N^{-1},\mathcal{F},\|\cdot\|_\infty)\geq 3$. Then,*

$$\mathbb{E}_{\mathcal{D}}\|f^\circ-\hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_X)}\leq C\left(\inf_{f\in\mathcal{F}}\|f-f^\circ\|^2_{\mathcal{L}^2(\mathcal{P}_X)}+\log\mathcal{N}\left(\frac{1}{N},\mathcal{F},\|\cdot\|_\infty\right)\frac{\tilde{F}^2}{N}\right),$$

where $C>0$ is a universal constant, $\tilde{F}:=\frac{R_{\mathcal{F}}}{\sigma}\vee\frac{\|f^\circ\|_\infty}{\sigma}\vee\frac{1}{2}$ and $R_{\mathcal{F}}:=\sup\{\|f\|_\infty\mid f\in\mathcal{F}\}$.

Proof. Basically, we convert our problem setting so that it fits to the assumptions of Lemma 4 of Schmidt-Hieber (2017) and apply the lemma to it. For $f:[-1,1]^D\to[-\sigma\tilde{F},\sigma\tilde{F}]$, we define $A[f]:[0,1]^D\to[0,2\tilde{F}]$ by $A[f](x'):=\frac{1}{\sigma}f(2x'-1)+\tilde{F}$. Let $\hat{f}_1$ be the (non-clipped) ERM etimator of $\mathcal{F}$. We define $X':=\frac{1}{2}(X+1)$, $f'^\circ:=A[f^\circ]$, $Y':=f'^\circ(X)+\xi'$, $\mathcal{F}':=\{A[f]\mid f\in\mathcal{F}\}$, $\hat{f}'_1:=A[\hat{f}_1]$, and $\mathcal{D}':=((x'_n,y'_n))_{n\in[N]}$ where $x'_n:=\frac{1}{2}(x_n+1)$ and $y'_n:=f'^\circ(x'_n)+\frac{1}{\sigma}(y_n-f^\circ(x_n))$. Then, the probability that $\mathcal{D}'$ is drawn from $\mathcal{P}'^{\otimes N}$ is same as the probability that $\mathcal{D}$ is drawn from $\mathcal{P}^{\otimes N}$ where $\mathcal{P}'$ is the joint distribution of (X',Y'). Also, we can show that $\hat{f}'$ is the ERM estimator of the regression problem $Y'=f'^\circ+\xi'$ using the dataset $\mathcal{D}'$: $\hat{f}'_1\in\arg\min_{f'\in\mathcal{F}'}\hat{\mathcal{R}}_{\mathcal{D}'}(f')$. We apply the Lemma 4 of Schmidt-Hieber (2017) with $n\leftarrow N$, $d\leftarrow D$, $\varepsilon\leftarrow 1$, $\delta\leftarrow\frac{1}{N}$, $\Delta_n\leftarrow 0$, $\mathcal{F}'\leftarrow\mathcal{F}$, $F\leftarrow 2\tilde{F}$, $\hat{f}\leftarrow\hat{f}'_1$ and use the fact that the estimation error of the clipped ERM estimator is no worse than that of the ERM estimator, that is, $\|f^\circ-\hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_X)}\leq\|f^\circ-\hat{f}_1\|^2_{\mathcal{L}^2(\mathcal{P}_X)}$ to conclude. □

Proof of Theorem 6. By Lemma 4, we have $\log\mathcal{N}:=\log\mathcal{N}(N^{-1},\mathcal{F}^{(\mathrm{CNN})},\|\cdot\|_\infty)\leq\Lambda_2\log(2B\Lambda_1N)$, where $B=B^{(\mathrm{conv})}\vee B^{(\mathrm{fc})}$. Therefore, by Lemma 5,

$$\begin{aligned}\|f^\circ-\hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_X)}&\leq C_0\left(\inf_{f\in\mathcal{F}}\|f-f^\circ\|^2_{\mathcal{L}^2(\mathcal{P}_X)}+\log\mathcal{N}\frac{\tilde{F}^2}{N}\right)\\&\leq C_1\left(\inf_{f\in\mathcal{F}^{(\mathrm{FNN})}}\|f-f^\circ\|^2_\infty+\frac{\tilde{F}^2}{N}\Lambda_2\log(2B\Lambda_1N)\right),\end{aligned}$$

where $C_0,C_1>0$ are universal constants. We used in the last inequality the fact $\|\mathrm{clip}[f]-f^\circ\|_{\mathcal{L}^2(\mathcal{P}_X)}\leq\|\mathrm{clip}[f]-f^\circ\|_\infty\leq\|f-f^\circ\|_\infty$ any $f\in\mathcal{F}^{(\mathrm{CNN})}$ and the assumption $\mathcal{F}^{(\mathrm{FNN})}\subset\mathcal{F}^{(\mathrm{CNN})}$. □

As discussed in the beginning of this section, Theorem 2 is the special case of Theorem 6.

Proof of Corollay 1. We only care the order with respect to N in the O-notation. Set $M=\lfloor N^\alpha\rfloor$ for $\alpha>0$. Using the assumptions of the corollary, the estimation error is

$$\|f^\circ-\hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_x)}=\tilde{O}\left(\max\left(N^{-2\alpha\gamma_1},N^{\alpha\gamma_2-1}\right)\right)$$

by Theorem 2. The order of the right hand side with respect to N is minimized when $\alpha=\frac{1}{2\gamma_1+\gamma_2}$. By substituting α, we can derive Corollary 1. □

E. Proofs of Corollary 2 and Corollary 3

By Theorem 2 of (Klusowski & Barron, 2018), for each $M\in\mathbb{N}_+$, there exists

$$f^{(\mathrm{FNN})}:=\frac{1}{M}\sum_{m=1}^M b_m(a_m^\top x-t_m)_+=\sum_{m=1}^M b_m\left(\frac{a_m^\top}{M}x-\frac{t_m}{M}\right)_+$$

with $|b_m|\leq 1$, $\|a_m\|_1=1$, and $|t_m|\leq 1$ such that $\|f^\circ-f^{(\mathrm{FNN})}\|_\infty\leq Cv_{f^\circ}\sqrt{\log M+D}M^{-\frac{1}{2}-\frac{1}{D}}$ where $v_{f^\circ}:=\int_{\mathbb{R}^D}\|w\|_2^s\,|\mathcal{F}[f^\circ](w)|\,\mathrm{d}w$ and $C>0$ is a universal constant. We set $L_m\leftarrow 1$, $D_m^{(1)}\leftarrow 1$, $B^{(\mathrm{bs})}\leftarrow\frac{1}{M}$, $B^{(\mathrm{fin})}\leftarrow 1$ ($m\in[M]$) in the Theorem 5, then, we have $f^{(\mathrm{FNN})}\in\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D}_1,B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$. By applying Theorem 5, there exists a CNN

$f^{(\mathrm{CNN})} \in \mathcal{F}^{(\mathrm{CNN})}_{\boldsymbol{C},\boldsymbol{K},B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$ such that $f^{(\mathrm{FNN})} = f^{(\mathrm{CNN})}$. Here, $\boldsymbol{C} = (C^{(1)}_m)_m$ with $C^{(1)}_m = 4$, $\boldsymbol{K} = (K^{(1)}_m)_m$ with $K^{(1)}_m = K$, $B^{(\mathrm{conv})} = \frac{1}{M}$, and $B^{(\mathrm{fc})} = M$. This proves Corollary 2.

With these evaluations, we have $\Lambda_1 = O(M^3)$ (note that since $B^{(\mathrm{conv})} = \frac{1}{M}$, we have $\prod_{m=0}^{M}(1+\rho_m) = O(1)$). In addition, $B^{(\mathrm{conv})}$ is $O(1)$ and $B^{(\mathrm{fc})}$ is $O(M)$. Therefore, we have $\log \Lambda_1 B = \tilde{O}(1)$. Since $\Lambda_2 = O(M)$, we can use Corollary 1 with $\gamma_1 = \frac{1}{2} + \frac{1}{D}$, $\gamma_2 = 1$. □

F. Proofs of Corollary 4 and Corollary 5

We first prove the scaling property of the FNN class.

Lemma 6. *Let $M \in \mathbb{N}_+$, $L_m \in \mathbb{N}_+$, and $D^{(l)}_m \in \mathbb{N}_+$ for $m \in [M]$ and $l \in [L_m]$. Let $B^{(\mathrm{bs})}, B^{(\mathrm{fin})} > 0$. Then, for any $k \geq 1$, we have $\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},B^{(\mathrm{bs})},B^{(\mathrm{fin})}} \subset \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},k^{-1}B^{(\mathrm{bs})},k^L B^{(\mathrm{fin})}}$ where $L := \max_{m\in[M]} L_m$ is the maximum depth of the blocks.*

Proof. Let $\boldsymbol{\theta} = ((W^{(l)}_m)_{m,l}, (b^{(l)}_m)_{m,l}, (w_m)_m, b)$ be the parameter of an FNN and suppose that $\mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}} \in \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},B^{(\mathrm{bs})},B^{(\mathrm{fin})}}$. We define $\boldsymbol{\theta}' := ((W'^{(l)}_m)_{m,l}, (b'^{(l)}_m)_{m,l}, (w'_m)_m, b')$ by

$$W'^{(l)}_m := k^{-\frac{L}{L_m}} W^{(l)}_m, \quad b'^{(l)}_m := k^{-l\frac{L}{L_m}} b^{(l)}_m, \quad w'_m := k^L w_m, \quad b' := b.$$

Since $k \geq 1$, we have $\mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}'} \in \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},k^{-1}B^{(\mathrm{bs})},k^L B^{(\mathrm{fin})}}$. Also, by the homogeneous property of the ReLU function (i.e., $\mathrm{ReLU}(ax) = a\mathrm{ReLU}(x)$ for $a > 0$), we have $\mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}} = \mathrm{FNN}^{\mathrm{ReLU}}_{\boldsymbol{\theta}'}$. □

Next, we prove the existence of a block-sparse FNN with constant-width blocks that optimally approximates a given β-Hölder function. It is almost same as the proof appeared in Schmidt-Hieber (2017). However, we need to construct the FNN so that it has a block-sparse structure.

Lemma 7 (cf. Schmidt-Hieber (2017) Theorem 5)**.** *Let $\beta > 0$, $M \in \mathbb{N}_+$ and $f^\circ : [-1,1]^D \to \mathbb{R}$ be a β-Hölder function. Then, there exists $D' = O(1)$, $L' = O(\log M)$, and a block-sparse FNN $f^{(\mathrm{FNN})} \in \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},1,2M\|f^\circ\|_\beta}$ such that $\|f^\circ - f^{(\mathrm{FNN})}\|_\infty = \tilde{O}(M^{-\frac{\beta}{D}})$. Here, we set $L_m := L'$ and $D^{(l)}_m := D'$ for all $m \in [M]$ and $l \in [L_m]$ and define $\boldsymbol{D} := (D^{(l)}_m)_{m,l}$.*

Proof. First, we prove the lemma when the domain of f° is $[0,1]^D$. Let M' be the largest interger satisfying $(M'+1)^D \leq M$. Let $\Gamma(M') = \left(\frac{\mathbb{Z}}{M'}\right)^D \cap [0,1]^D = \{\frac{m'}{M'} \mid m' \in \{0,\dots,M'\}^D\}$ be the set of lattice points in $[0,1]^D$[2]. Note that the cardinality of $\Gamma(M')$ is $(M'+1)^D$. Let $P^\beta_a f^\circ$ be the Taylor expansion of f° up to order $\lfloor\beta\rfloor$ at $a \in [0,1]^D$:

$$(P^\beta_a f^\circ)(x) = \sum_{0\leq|\alpha|<\beta} \frac{(\partial^\alpha f^\circ)(a)}{\alpha!}(x-a)^\alpha.$$

For $a \in [0,1]^D$, we define a hat-shaped function $H_a : [0,1]^D \to [0,1]$ by

$$H_a(x) := \prod_{j=1}^{D} (M'^{-1} - |x_j - a_j|_+).$$

Note that we have $\sum_{a\in\Gamma(M')} H_a(x) = 1$, i.e., they are a partition of unity. Let $P^\beta f^\circ$ be the weighted sum of the Taylor expansions at lattice points of $\Gamma(M')$:

$$(P^\beta f^\circ)(x) := M'^D \sum_{a\in D(M')} (P^\beta_a f^\circ)(x) H_a(x).$$

By Lemma B.1 of Schmidt-Hieber (2017), we have

$$\|P^\beta f^\circ - f^\circ\|_\infty \leq \|f^\circ\|_\beta M'^{-\beta}.$$

[2] Schmidt-Hieber (2017) used $\boldsymbol{D}(M')$ to denote this set of lattice points. We used different character to avoid notational conflict.

Let m be an interger specified later and set $L^* := (m+5)\lceil \log_2 D \rceil$. By the proof of Lemma B.2 of Schmidt-Hieber (2017), for any $a \in \Gamma(M')$, there exists an FNN $\mathrm{Hat}_a : [0,1]^D \to [0,1]$ whose depth and width are at most $2 + L^*$ and $6D$, respectively and whose parameters have sup-norm 1, such that

$$\|\mathrm{Hat}_a - H_a\|_\infty \le 3^D 2^{-m}.$$

Next, let $B := 2\|f^\circ\|_\beta$ and $C_{D,\beta}$ be the number of distinct D-variate monomials of degree up to $\lfloor \beta \rfloor$. By the equation (7.11) of Schmidt-Hieber (2017), for any $a \in \Gamma(M)$, there exists an FNN $Q_a : [0,1]^D \to [0,1]$ [3] whose depth and width are $1 + L^*$ and $6DC_{D,\beta}$ respectively and whose parameters have sup-norm 1, such that

$$\left\| Q_a - \left(\frac{P_a^\beta f^\circ}{B} + \frac{1}{2} \right) \right\|_\infty \le 3^D 2^{-m}.$$

Thirdly, by Lemma A.2 of (Schmidt-Hieber, 2017), there exists an FNN $\mathrm{Mult} : [0,1]^2 \to [0,1]$, whose depth and width are $m+4$ and 6, respectively and whose parameters have sup-norm 1 such that

$$|\mathrm{Mult}(x,y) - xy| \le 2^{-m}$$

for any $x, y \in [0,1]$. For each $a \in \Gamma(M')$, we combine Hat_a and Q_a using Mult and constitute a block of the block-sparse FNN corresponding to $a \in \Gamma(M)$ by $\mathrm{FC}_a := \mathrm{Mult}(Q_a(\cdot), \mathrm{Hat}_a(\cdot))$. Then, we have

$$\left\| \mathrm{FC}_a - \left(\frac{P_a^\beta f^\circ}{B} + \frac{1}{2} \right) H_a \right\|_\infty \le 2^{-m} + 3^D 2^{-m} + 3^D 2^{-m} \le 3^{D+1} 2^{-m}.$$

We define $f^{(\mathrm{FNN})}(x) := \sum_{a \in \Gamma(M)} (BM'^D \mathrm{FC}_a(x)) - \frac{B}{2}$. By construction, $f^{(\mathrm{FNN})}$ is a block-sparse FNN with $(M'+1)^D (\le M)$ blocks each of which has depth and width at most $L' := 2 + L^* + (m+4)$ and $D' := 6(C_{D,\beta}+1)D$, respectively. The norms of the block-sparse part and the finally fully-connected layer are 1 and $BM'^D (\le BM)$, respectively. In addition, we have

$$\begin{aligned}
&|f^{(\mathrm{FNN})}(x) - (P^\beta f^\circ)(x)| \\
&\le \sum_{a \in \Gamma(M)} BM'^D \left| \mathrm{FC}_a(x) - \left(\frac{(P_a^\beta f^\circ)(x)}{B} + \frac{1}{2} \right) H_a(x) \right| + \frac{B}{2} \left| 1 - M'^D \sum_{a \in \Gamma(M')} H_a(x) \right| \\
&\le (M'+1)^D \times BM'^D 3^{D+1} 2^{-m} \\
&\le 3^{D+1} 2^{-m} BM^2
\end{aligned}$$

for any $x \in [0,1]^D$. Therefore,

$$\begin{aligned}
|f^{(\mathrm{FNN})}(x) - f^\circ(x)| &\le |f^{(\mathrm{FNN})} - (P^\beta f^\circ)(x)| + |(P^\beta f^\circ)(x) - f^\circ(x)| \\
&\le 3^{D+1} 2^{-m} BM^2 + \|f^\circ\|_\beta M'^{-\beta} \\
&\le 2 \cdot 3^{D+1} 2^{-m} \|f^\circ\|_\beta M^2 + \|f^\circ\|_\beta M^{-\frac{\beta}{D}}.
\end{aligned}$$

We set $m = \lceil \log_2 M^{2+\frac{\beta}{D}} \rceil$, then, we have $L' = O(\log M)$, $D' = O(1)$, and

$$\|f^{(\mathrm{FNN})} - f^\circ\| \le \|f^\circ\|_\beta (2 \cdot 3^{D+1} + 2^\beta) M^{-\frac{\beta}{D}}.$$

By the defnition of $f^{(\mathrm{FNN})}$ we have $f^{(\mathrm{FNN})} \in \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D}, 1, 2\|f^\circ\|_\beta M}$.

When the domain of f° is $[-1,1]^D$, we should add the function $x \mapsto \frac{1}{2}(x+1) = \frac{1}{2}(x+1)_+ - \frac{1}{2}(-x-1)_+$ as a first layer of each block to fit the range into $[0,1]^D$. Specifically, suppose the first layer of m-th block in $f^{(\mathrm{FNN})}$ is $x \mapsto \mathrm{ReLU}(Wx - b)$, then the first two layers become $x \mapsto \mathrm{ReLU}(\begin{bmatrix} \frac{1}{2}(x+1) & -\frac{1}{2}(x+1) \end{bmatrix})$ and $\begin{bmatrix} y_1 & y_2 \end{bmatrix} \mapsto \mathrm{ReLU}(Wy_1 - Wy_2 - b)$, respectively. Since this transformation does not change the maximum sup norm of parameters in the block-sparse and the order of L' and D', the resulting FNN still belongs to $\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D}, 1, 2\|f^\circ\|M}$. □

[3] We prepare Q_a for each $a \in \Gamma(M)$ as opposed to the original proof of (Schmidt-Hieber, 2017), in which Q_a's shared the layers the except the final one and were collectively denoted by Q_1.

Proofs of Corollary 4 and Corollary 5. In this proof, we only care the dependence on M in the O-notation. Let $\tilde{M} := 2\|f^\circ\|_\beta M$. By Lemma 7, there exists $f^{(\mathrm{FNN})} \in \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},1,\tilde{M}}$ such that $\|f^{(\mathrm{FNN})} - f^\circ\|_\infty = O(M^{-\frac{\beta}{D}})$ (L', D', and $\boldsymbol{D}$ as in Lemma 7). Let $k := 16D'K(M^{\frac{1}{L'}} \wedge 1)^{-1} = 16D'K(e^{\frac{1}{C'}} \wedge 1)^{-1} \geq 1$ where C' is a constant such that $L' = C' \log M$. Using Lemma 6, there exists $\tilde{f}^{(\mathrm{FNN})} \in \mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},k^{-1},k^{L'}\tilde{M}}$ such that $\tilde{f}^{(\mathrm{FNN})} = f^{(\mathrm{FNN})}$. We apply Theorem 5 to $\mathcal{F}^{(\mathrm{FNN})}_{\boldsymbol{D},k^{-1},k^{L'}\tilde{M}}$ and find $f^{(\mathrm{CNN})} \in \mathcal{F}^{(\mathrm{CNN})}_{\boldsymbol{C},\boldsymbol{K},B^{(\mathrm{conv})},B^{(\mathrm{fc})}}$ such that $L \leq M(L'+L_0)$, $\boldsymbol{C} := (C^{(l)}_m)_{m\in[M],l\in[L_m]}$ with $C^{(l)}_m \leq 4D'$, $\boldsymbol{K} := (K^{(l)}_m)_{m\in[M],l\in[L_m]}$ with $K^{(l)}_m \leq K$, $B^{(\mathrm{conv})} = k^{-1}$, $B^{(\mathrm{fc})} = k^{L'}(k\vee 1)\tilde{M} = k^{L'+1}\tilde{M}$, and $f^{(\mathrm{CNN})} = \tilde{f}^{(\mathrm{FNN})}$. This proves Corollary 4 (note that by definition, we have $B^{(\mathrm{conv})} = k^{-1} = O(1)$ and $\log B^{(\mathrm{fc})} = (L'+1)k + \log(\tilde{M}) = O(\log M)$).

By the definition of k and the bound on $C^{(l)}_m$ and $K^{(l)}_m$, we have $C^{(l-1)}_m K^{(l)}_m k^{-1} \leq \frac{1}{4}M^{-\frac{1}{L'}}$. Therefore, we have $\rho_m \leq \prod_{l=1}^{L'}(C^{(l-1)}_m K^{(l)}_m k^{-1}) \leq M^{-1}$ and hence $\prod_{m=0}^{M}(1+\rho_m) = O(1)$. Since $C^{(l-1)}_m K^{(l)}_m k^{-1} \leq \frac{1}{2}$ for sufficiently large M, we have $\rho^+_m = 1$ for sufficiently large M. In addition, we have $\log(B^{(\mathrm{conv})} \vee B^{(\mathrm{fc})}) = \tilde{O}(1)$. Combining them, we have $\log \Lambda_1 = O(1)$ and hence $\log \Lambda_1(B^{(\mathrm{conv})} \vee B^{(\mathrm{fc})}) = \tilde{O}(1)$. For Λ_2, we can bound it by $\Lambda_2 = O(M \log M)$ using bounds for $C^{(l)}_m$, $K^{(l)}_m$ and L'. Therefore, we can apply Corollary 2 with $\gamma_1 = \frac{\beta}{D}$, $\gamma_2 = 1$ and obtain the desired estimation error. Since we have $M = O(N^{\frac{1}{2\gamma_1+\gamma_2}})$ by the proof of Corollary 1, we can derive the bounds for L_m with respect to N. □

G. Proofs of Theorem 3 and Theorem 4

Lemma 8. *Let $L, L', C', K' \in \mathbb{N}_+$ and $B > 0$. Suppose we can realize $f + \mathrm{id} : \mathbb{R}^{D\times C'} \to \mathbb{R}^{D\times C'}$ with a residual block with an identity connection whose depth, channel size, and filter size are L', C', and K', respectively and whose parameter norm is bounded by B. Let $S_0 = \lceil \frac{L'}{L} \rceil$. Then, there exist $S = 2S_0 - 1$ functions $\tilde{f}_1, \dots, \tilde{f}_S : \mathbb{R}^{D\times 3C'} \to \mathbb{R}^{D\times 3C'}$ and S masks $z_1, \dots, z_S \in \{0,1\}^{3C'}$, such that f_s is realizable by a residual block whose depth, channel size, filter size, and parameter norm bound are L, $3C'$, K', and B, respectively and $\tilde{f} := (\tilde{f}_S + J_S) \circ \cdots \circ (\tilde{f}_1 + J_1) : \mathbb{R}^{D\times 3C'} \to \mathbb{R}^{D\times 3C'}$ satisfies $\tilde{f}(\begin{bmatrix} x & 0 & 0 \end{bmatrix}) = \begin{bmatrix} f(x) & 0 & 0 \end{bmatrix}$. Here J_s is a channel-wise mask operation made from z_s.*

Proof. We divide the residual block representing f into S_0 CNNs with depth at most L and denote them sequentially by $g_1, \dots, g_{S_0}$ so that $f = g_{S_0} \circ \cdots \circ g_1$. We define $\tilde{g}_s : \mathbb{R}^{D\times 3C'} \to \mathbb{R}^{D\times 3C'}$ ($s \in [S_0]$) from g_s by

$$\tilde{g}_s([x_1\ x_2\ x_3]) = \begin{cases} [0\ y_1\ 0] \text{ (if } s = 1) \\ [0\ y_3\ 0] \text{ (if } s \neq 1, S_0 \text{ and odd)} \\ [0\ 0\ y_2] \text{ (if } s \neq 1, S_0 \text{ and even)} \\ [y_3\ 0\ 0] \text{ (if } s = S_0 \text{ and odd)} \\ [y_2\ 0\ 0] \text{ (if } s = S_0 \text{ and even)} \end{cases},$$

where $y_i = g_s(x_i)$ ($i = 1, 2, 3$). Note that we can construct $\tilde{g}_s$ by a residual block with depth L, channel size $3C'$, filter size K', and parameter norm B. Next, we define u_s ($s \in [S_0 - 1]$) by

$$u_s = \begin{cases} \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top & \text{(if } s\text{: odd)} \\ \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^\top & \text{(if } s\text{: even)} \end{cases}$$

Then, we define $\tilde{f} := (\tilde{g}_{S_0} + \mathrm{id}) \circ (0 + J'_{S_0-1}) \circ (\tilde{g}_{S_0-1} + \mathrm{id}) \circ (0 + J'_1) \circ (\tilde{f}_1 + \mathrm{id})$ where J'_s is a channel-wise mask constructed from u_s and $0 : \mathbb{R}^{D\times 3C'} \to \mathbb{R}^{D\times 3C'}$ is a constant zero function, which is obviously representable by a residual block. By definition, $\tilde{f}$ is realizable by S residual blocks with channel-wise masking identity connections and satisfy the conditions on the depth, channel size, filter size, and norm bound. □

Proof of Theorem 3. The first part of the proof is same as that of Corollary 4, except that we define k using L instead of L': $k = 16D'K(M^{\frac{1}{L}} \wedge 1)^{-1}$. Here, D' is a constant satisfying $D' = O(1)$ as a function of M. Then, there exists a CNN $\tilde{f}^{(\mathrm{CNN})} \in \mathcal{F}^{(\mathrm{CNN})}_{M,L',C',K',B^{(\mathrm{conv})},B^{(\mathrm{fin})}}$ such that $\|\tilde{f}^{(\mathrm{CNN})} - f^\circ\| = O(M^{-\frac{\beta}{D}})$. Parameter of the set of CNNs satisfy $L' = O(\log M)$ $C' \leq 4D'$, $K' \leq K$, $B^{(\mathrm{conv})} = k^{-1}$, and $B^{(\mathrm{fc})} = 2\|f^\circ\|_\beta k^{L'} M$. We apply Lemma 8 to each residual block of $\tilde{f}^{(\mathrm{CNN})}$. Then, there exists $f^{(\mathrm{CNN})} \in \mathcal{G}_{\tilde{M},L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fin})}}$ such that $f^{(\mathrm{CNN})} = \tilde{f}^{(\mathrm{CNN})}$ and $\tilde{M} = M\lceil \frac{L'}{L} \rceil$, $C \leq 3C'$, $K' \leq K$, $B^{(\mathrm{conv})} = k^{-1}$, and $B^{(\mathrm{fin})} = 2\|f^\circ\|_\beta k^{L'+1} M$. □

Before going to the proof of Theorem 4, we first note that the definitions of Λ_1 and Λ_2 in Theorem 2 are valid even if we replace $\mathcal{F}^{(\mathrm{CNN})}_{\tilde{M},L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fin})}}$ with $\mathcal{G} = \mathcal{G}_{\tilde{M},L,C,K,B^{(\mathrm{conv})},B^{(\mathrm{fin})}}$.

Lemma 9. *Let $\tilde{M}, L, C, K \in \mathbb{N}_+$ and $B^{(\mathrm{conv})}, B^{(\mathrm{fin})}, \varepsilon > 0$. Set $B = B^{(\mathrm{conv})} \vee B^{(\mathrm{fin})}$. Then, the covering number of $\mathcal{G}$ with respect to the sup-norm $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_\infty)$ is bounded by $(2B\Lambda_1\varepsilon^{-1})^{\Lambda_2} \cdot 2^{C\tilde{M}L}$, where $\Lambda_1 = \Lambda_1(\mathcal{G})$ and $\Lambda_2 = \Lambda_2(\mathcal{G})$ are ones defined in Theorem 2, except that $\mathcal{F}^{(\mathrm{CNN})}$ is replaced with $\mathcal{G}$.*

Proof. First we note that we can apply same inequalities in Section D.2.1 – D.2.3 and Proposition 11 to CNNs in $\mathcal{G}$. Therefore, if two masked CNNs $f, g \in \mathcal{G}$ have same masking patterns in identity connections and distance of each pair of corresponding parameters in residual blocks is at most ε, then, we can show $\|f - g\|_\infty \le \Lambda_1\varepsilon$ in the same way as Lemma 3. Therefore, by the same argument of Lemma 4, the covering number of the subset of $\mathcal{G}$ consisting of CNNs with a specific masking pattern is bounded by $(2B\Lambda_1\varepsilon^{-1})^{\Lambda_2}$. Since each CNN in $\mathcal{G}$ has $C\tilde{M}L$ parameters in identity connections which take values in $\{0, 1\}$, there are $2^{C\tilde{M}L}$ masking patterns. Therefore, we have $\mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_\infty) \le (2B\Lambda_1\varepsilon^{-1})^{\Lambda_2} \cdot 2^{C\tilde{M}L}$. □

The strategy for the proof of Theorem 4 is almost same as the proofs for Theorem 6 and Corollary 5, except that we should replace $\Lambda_2 \log(2B\Lambda_1 N)$ in (3) with $\Lambda_2 \log(2B\Lambda_1 N) + C\tilde{M}L \log 2$ (and Λ_1 and Λ_2 are defined via $\mathcal{G}$ instead of $\mathcal{F}^{(\mathrm{CNN})}$). However, the second term is at most as same order (upto logarithmic factors) as the first one in our situation. Therefore, we can derive the same estimation error rate.

Proof of Theorem 4. Take $\mathcal{G}$ as in the proof of Theorem 3. Let $\log \mathcal{N} := \log \mathcal{N}(N^{-1}, \mathcal{G}, \|\cdot\|_\infty)$. By Lemma 5, we have

$$\|f^\circ - \hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_X)} \le C_0 \left(\inf_{f \in \mathcal{F}^{(\mathrm{FNN})}} \|f - f^\circ\|^2_\infty + \frac{\tilde{F}^2}{N} \left(\Lambda_2 \log(2B\Lambda_1 N) + C\tilde{M}L \log 2 \right) \right),$$

where $C_0 > 0$ is a universal constant. The first term in the outer-most parenthesis is $O(M^{-\frac{\beta}{D}})$ by Lemma 7. We will evaluate the order of the second term. First, we have $\Lambda_2 = O(\tilde{M}) = \tilde{O}(M)$ by the definition of Λ_2. By the definition of k, we have $\rho \le M^{-1}$ and $\rho^+ = 1$ for sufficiently large M therefore, $\varrho = O(1)$ and $\varrho^+ = O(M)$ for sufficiently large M. Again, by the definition of k, we have $B^{(\mathrm{conv})} = O(1)$ and $B^{(\mathrm{fc})} = O(M)$. Therefore, we have $\Lambda_1 = O(M^3)$ and $B = O(M)$ and hence $\Lambda_2 \log(2B\Lambda_1 N) = \tilde{O}(MN)$. On the other hand, since $C = O(1), \tilde{M} = \tilde{O}(M), L = O(1)$, we have $C\tilde{M}L \log 2 = \tilde{O}(M)$.

Therefore, by setting $M = \lfloor N^\alpha \rfloor$ for $\alpha > 0$, the estimation error is

$$\|f^\circ - \hat{f}\|^2_{\mathcal{L}^2(\mathcal{P}_x)} = \tilde{O}\left(\max\left(N^{-2\alpha\gamma_1}, N^{\alpha\gamma_2 - 1} \right) \right),$$

where $\gamma_1 = \frac{\beta}{D}$ and $\gamma_2 = 1$. The order of the right hand side with respect to N is minimized when $\alpha = \frac{1}{2\gamma_1 + \gamma_2}$. By substituting α, we can derive the theorem. □

H. One-sided padding vs. Equal-padding

In this paper, we adopted one-sided padding, which is not used so often practically, in order to make proofs simple. However, with slight modifications, all statements are true for equally-padded convolutions, a widely employed padding style which adds (approximately) same numbers of zeros to both ends of an input signal, with the exception that the filter size K is restricted to $K \le \lfloor \frac{D}{2} \rfloor$ instead of $K \le D$.

I. Difference between Original ResNet and Ours

There are several differences between the CNN in this paper and the original ResNet (He et al., 2016), aside from the number of layers. The most critical one is that our CNN does not have pooling nor Batch Normalization layers (Ioffe & Szegedy, 2015). We will consider a scaling scheme simpler than Batch Normalization to derive optimality of CNNs with constant-depth residual blocks (see Definition 5). It is left for future research whether our result can extend to the ResNet-type CNNs with pooling or other scaling layers such as Batch Normalization.

References

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.

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.

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.

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.

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

Shang, W., Sohn, K., Almeida, D., and Lee, H. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2217–2225. PMLR, 2016.