

Mathematical Techniques in the Approximation Theory that are Rooted in Neural Networks - Petersen

Ko, Suh, Huo

Georgia Tech

Summer of 2020

Table of Contents

1 Basics of Neural Networks

- Definition
- Complexity of Neural Network
- Operations of Neural Network

2 Approximation with Neural Networks

- Mathematical Problem
- Exemplary Results
- Characterization of Classifier Functions
- Approximation of Smooth Functions
- Approximation of Horizon Functions
- Approximation of Piecewise Constant Functions
- Approximation of Piecewise Smooth Functions

Definition of Neural Network

We define neural network as a sequence of matrix-vector tuples:

- $d, L \in \mathbb{N}$
- $N_0 = d, N_1, \dots, N_L \in \mathbb{N}, l \in \{1, \dots, L\}$
- A_1, A_2, \dots, A_L such that $A_l \in \mathbb{R}^{N_l \times N_{l+1}}$
- b_1, b_2, \dots, b_L such that $b_l \in \mathbb{R}^{N_l}$
- $\rho : \mathbb{R} \rightarrow \mathbb{R}$

Then, a neural network is

$$\Phi = ((A_1, b_1), (A_2, b_2), \dots, (A_L, b_L))$$

such that its realization $R(\Phi) : \mathbb{R}^d \rightarrow \mathbb{R}^{N_L}$ is a function such that $R(\Phi)(x) = x_L$ where

$$X_0 := X$$

$$x_l = \varrho(A_l x_{l-1} + b_l)$$

$$x_l = A_l x_{l-1} + b_l$$

Operations of Neural Network

Realization of Identity

Let ρ be ReLU, let $d \in \mathbb{N}$ and define

$$\Phi_d^{ld} := ((A_1, b_1), (A_2, b_2))$$

where

$$A_1 := \begin{pmatrix} \text{Id}_{\mathbb{R}^d} \\ -\text{Id}_{\mathbb{R}^d} \end{pmatrix} \quad b_1 := 0, \quad A_2 := (\text{Id}_{\mathbb{R}^d} \quad -\text{Id}_{\mathbb{R}^d}) \quad b_2 := 0$$

Then, $R_\rho(\Phi_d^{\text{Id}}) = \text{Id}_{\mathbb{R}^d}$

As a generalization, one can arbitrarily adjust the number of layers by adding arbitrary number of matrix-vector pairs $(\text{Id}_{\mathbb{R}^{2d}}, 0)$ in between (A_1, b_1) and (A_2, b_2) above. This yields $\Phi_{d,L}^{\text{Id}}$ with L layers. Trivially, in case one wants $L = 1$, $\Phi_{d,1}^{\text{Id}} := ((\text{Id}_{\mathbb{R}^d}, 0))$ suffices.

Operations of Neural Networks

Concatenation

Let $L_1, L_2 \in \mathbb{N}$ and let $\Phi^1 = ((A_1^1, b_1^1), \dots, (A_{L_1}^1, b_{L_1}^1))$ and $\Phi^2 = ((A_1^2, b_1^2), \dots, (A_{L_2}^2, b_{L_2}^2))$ be two neural networks such that the input layer of Φ^1 has the same dimension as the output layer of Φ^2 . Then, $\Phi^1 \bullet \Phi^2$ denotes the following $L_1 + L_2 - 1$ layer network:

$$\Phi^1 \bullet \Phi^2 := ((A_1^2, b_1^2), \dots, (A_{L_2-1}^2, b_{L_2-1}^2), (A_1^1 A_{L_2}^2, A_1^1 b_{L_2}^2 + b_1^1), \\ (A_2^1, b_2^1), \dots, (A_{L_1}^1, b_{L_1}^1))$$

We call $\Phi^1 \bullet \Phi^2$ the concatenation of Φ^1 and Φ^2

Sparse Concatenation

Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the ReLU, let $L_1, L_2 \in \mathbb{N}$, and let Φ^1, Φ^2 be two neural networks as above. Let Φ_d^{Id} be as in the previous slide. Then, the sparse concatenation of Φ^1 and Φ^2 is defined as

$$\Phi^1 \odot \Phi^2 := \Phi^1 \bullet \Phi_d^{\text{id}} \bullet \Phi^2$$

Operations of Neural Networks

Properties of sparse concatenation

It follows immediately from above definitions that

$$\begin{aligned} \Phi^1 \odot \Phi^2 = & \left((A_1^2, b_1^2), \dots, (A_{L_2-1}^2, b_{L_2-1}^2), \left(\begin{pmatrix} A_{L_2}^2 \\ -A_{L_2}^2 \end{pmatrix}, \begin{pmatrix} b_{L_2}^2 \\ -b_{L_2}^2 \end{pmatrix} \right), \right. \\ & \left. \left([A_1^1 \mid -A_1^1], b_1^1 \right), (A_2^1, b_2^1), \dots, (A_{L_1}^1, b_{L_1}^1) \right) \end{aligned}$$

Then, we can deduce the following properties:

- $L(\Phi^1 \odot \Phi^2) = L_1 + L_2$
- $R_\varrho(\Phi^1 \odot \Phi^2) = R_\varrho(\Phi^1) \circ R_\varrho(\Phi^2)$
- $M(\Phi^1 \odot \Phi^2) \leq 2M(\Phi^1) + 2M(\Phi^2)$

Using the fact $a + b \leq 2 \max\{a, b\}$ for $a, b \geq 0$, it follows inductively

$$M(\Phi^1 \odot \dots \odot \Phi^n) \leq 4^{n-1} \cdot \max \{M(\Phi^1), \dots, M(\Phi^n)\}$$

Operations of Neural Networks

Parallelization

Let $L \in \mathbb{N}$ and let $\Phi^1 = ((A_1^1, b_1^1), \dots, (A_L^1, b_L^1))$ and $\Phi^2 = ((A_1^2, b_1^2), \dots, (A_L^2, b_L^2))$ be two neural networks with L layers and with d -dimensional input. We define

$$P(\Phi^1, \Phi^2) := \left((\tilde{A}_1, \tilde{b}_1), \dots, (\tilde{A}_L, \tilde{b}_L) \right)$$

where

$$\tilde{A}_1 := \begin{pmatrix} A_1^1 \\ A_1^2 \end{pmatrix}, \tilde{b}_1 := \begin{pmatrix} b_1^1 \\ b_1^2 \end{pmatrix} \quad \text{and} \quad \tilde{A}_\ell := \begin{pmatrix} A_\ell^1 & 0 \\ 0 & A_\ell^2 \end{pmatrix}, \tilde{b}_\ell := \begin{pmatrix} b_\ell^1 \\ b_\ell^2 \end{pmatrix}$$

for $1 < \ell \leq L$. Then, $P(\Phi^1, \Phi^2)$ is a neural network with d -dimensional input and L layers, called the parallelization of Φ^1 and Φ^2 . One can check $M(P(\Phi^1, \Phi^2)) = M(\Phi^1) + M(\Phi^2)$ and

$$R_\varrho \left(P \left(\Phi^1, \Phi^2 \right) \right) (x) = \left(R_\varrho \left(\Phi^1 \right) (x), R_\varrho \left(\Phi^2 \right) (x) \right) \quad \text{for all } x \in \mathbb{R}^d$$

Parallelization

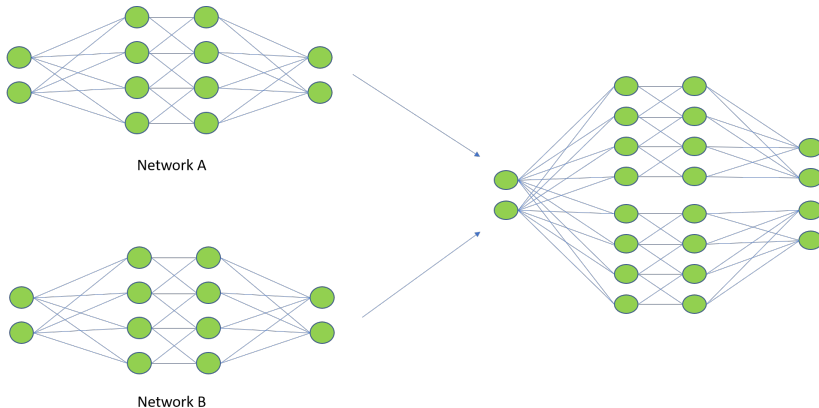


Figure 1: Parallelization.

Approximation with Neural Networks

Mathematical Problem

- Given a function $f \in \mathcal{C}$ where \mathcal{C} is some class of functions, how many weights, nodes, and layers does one need, and how much quantization do we require to approximate f to ε accuracy in some predefined metric?

Exemplary Results

Suppose $f \in \mathcal{C} \subset L^\infty(\Omega)$.

- Barron; 1993: If \mathcal{C} are bounded functions with one finite Fourier moment and ϱ is sigmoidal, then there exist NNs $(\Phi_n^\varrho)_{n \in \mathbb{N}}$ such that $\|f - \Phi_n^\varrho\|_{L^\infty} \lesssim N(\Phi_n^\varrho)^{-1} \rightarrow 0$
- Mhaskar; 1993: For $\mathcal{C} = C^s([0, 1]^d)$, ϱ a sigmoidal function of order $k \geq 2$, and $L(\Phi_n^\varrho) = L(k, d, s)$, there exist NNs $(\Phi_n^\varrho)_{n \in \mathbb{N}}$ such that $\|f - \Phi_n^\varrho\|_{L^\infty} \lesssim N(\Phi_n^\varrho)^{-s/d} \rightarrow 0$
- Yarotsky; 2017: For $\mathcal{C} = C^S([0, 1]^d)$ and $\varrho(x) = \max\{0, x\}$ (called ReLU) and $L(\Phi_n^\varrho) \asymp \log(n)$ we have that there exist NNs $(\Phi_n^\varrho)_{n \in \mathbb{N}}$ such that $\|f - \Phi_n^\varrho\|_{L^\infty} \lesssim W(\Phi_n^\varrho)^{-s/d} \rightarrow 0$

Characterization of Classifier Functions

Smooth functions

For a given $\beta \in (0, \infty)$ with $\beta = n + \sigma$ for $n \in \mathbb{N}_0, \sigma \in (0, 1]$, let $f \in C^n([-\frac{1}{2}, \frac{1}{2}]^r)$. Define

$$\|f\|_{C^{0,\beta}} := \max \left\{ \max_{|\alpha| \leq n} \|\partial^\alpha f\|_{\text{sup}}, \max_{|\alpha|=n} \text{Lip}_\sigma(\partial f) \right\} \in [0, \infty]$$

where

$$\text{Lip}_\sigma(g) := \sup_{x,y \in \Omega, x \neq y} \frac{|g(x) - g(y)|}{|x - y|^\sigma} \text{ for } g : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$$

Then, for $B > 0$, we define the following class of smooth functions:

$$\mathcal{F}_{\beta,d,B} := \{f \in C^n([-1/2, 1/2]^d) : \|f\|_{C^{0,\beta}} \leq B\}$$

Characterization of Classifier Functions

Horizon functions

We define the class of horizon functions as:

$$\mathcal{HF}_{\beta,d,B} := \{f \circ T \in L^\infty([-1/2, 1/2]^d) : f(x) = H(x_1 + \gamma(x_2, \dots, x_d), x_2, \dots, x_d), \gamma \in \mathcal{F}_{\beta,d-1,B}, T \in \Pi(d, \mathbb{R})\}$$

where $\Pi(d, \mathbb{R}) \subset GL(d, \mathbb{R})$ denotes the group of permutation matrices.

Piecewise constant functions

Let $r \in \mathbb{N}$, $d \in \mathbb{N}_{>2}$, and $\beta, B > 0$. Define

$$\mathcal{K}_{r,\beta,d,B} := \{K \subset [-1/2, 1/2]^d : \forall x \in [-1/2, 1/2]^d, \exists f_x \in \mathcal{HF}_{\beta,d,B} : \\ \chi_K = f_x \text{ on } [-1/2, 1/2]^d \cap \overline{B_{2^{-r}}}\|_{\ell^\infty}(x)\}$$

Then, the class of functions $\{\sum_{i \leq N} \chi_{K_i} : K_i \in \mathcal{K}_{r,\beta,d,B}, N \in \mathbb{N}\}$ characterizes piecewise constant functions.

Characterization of Classifier Functions

Piecewise smooth functions

Finally, we consider the following class of piecewise smooth functions:

$$\mathcal{E}_{r,\beta,d,B}^p := \{f = \chi_K \cdot g : g \in \mathcal{F}_{\beta',d,B} \text{ and } K \in \mathcal{K}_{r,\beta,d,B}\}$$

Note that unlike in the case of piecewise constant functions, the smoothness of the boundary surface, β , is allowed to differ from the smoothness of smooth regions, β' .

Approximation results

In the next set of slides, we study approximation capabilities of neural networks for the 4 classes of functions introduced above in the usual metric defined by L^p -norm in the domain $[-\frac{1}{2}, \frac{1}{2}]^d$.

Approximation of Smooth Functions (Theorem 3.1, P., Voigtlaender 2018)

For any $d \in \mathbb{N}$, and $\beta, B, p > 0$, there exist constants $s = s(d, \beta, B, p) \in \mathbb{N}$ and $c = c(d, \beta, B) > 0$ such that for any function $f \in \mathcal{F}_{\beta, d, B}$ and any $\varepsilon \in (0, 1/2)$, there is a neural network Φ_ε^f with at most $(2 + \lceil \log_2 \beta \rceil) \cdot (11 + \beta/d)$ layers, and at most $c \cdot \varepsilon^{-d/\beta}$ nonzero, (s, ε) -quantized weights such that

$$\left\| R_{\varrho} \left(\Phi_{\varepsilon}^f \right) - f \right\|_{L^p([-1/2, 1/2]^d)} < \varepsilon \quad \text{and} \quad \left\| R_{\varrho} \left(\Phi_{\varepsilon}^f \right) \right\|_{\text{sup}} \leq \lceil B \rceil$$

Key Proof Ideas

- Construct neural networks that can approximate multiplication, monomials, and polynomials up to arbitrary accuracy at the cost of some level of network complexity.
- Construct an equally spaced n -dimensional grid on the domain of f , $[-\frac{1}{2}, \frac{1}{2}]^d$.
- Consider Taylor approximation of f by $p_{i,\alpha}$ where $p_{i,\alpha}$ has x_i in one of the grid cubes as its base point.
- Construct a neural network whose number of outputs equals the number of grid cubes that can approximate any of $p_{i,\alpha}$ with one of its outputs.
- Construct a neural network with 1-dimensional output that approximates each output of the network from previous step restricted to the corresponding grid cubes.
- Modify the network from last step to ensure it has bounded outputs and make sure it satisfies all the stated properties regarding approximation accuracy, complexity, and number of layers.

Bounding the Outputs of Neural Networks (Lemma A.1)

There is a universal constant $c > 0$ such that the following holds:

For arbitrary $d, s, k \in \mathbb{N}$, $B > 0$, $\varepsilon \in (0, 1/2)$, and any neural network Ψ with d -dimensional input and k -dimensional output and with (s, ε) -quantized weights, there exists a neural network Φ with the same input/output dimensions as Ψ and with the following properties:

- $M(\Phi) \leq 2M(\Psi) + ck$, and $L(\Phi) \leq L(\Psi) + 2$
- All weights of Φ are (s_0, ε) -quantized, where $s_0 := \max\{\lceil \log_2(\lceil B \rceil) \rceil, s\}$
- $R_\rho(\Phi) = (\tau_B \times \cdots \times \tau_B) \circ R_\rho(\Psi)$, where the function

$$\tau_B : \mathbb{R} \rightarrow [-\lceil B \rceil, \lceil B \rceil], y \mapsto \text{sign}(y) \cdot \min\{|y|, \lceil B \rceil\}$$

is 1-Lipschitz and satisfies $\tau_B(y) = y$ for all $y \in \mathbb{R}$ with $|y| \leq [B]$

Approximation of Multiplication (Lemma A.3)

Let $\theta > 0$ be arbitrary. Then, for every $L \in \mathbb{N}$ with $L > (2\theta)^{-1}$ and each $M \geq 1$, there are constants $c = c(L, M, \theta) \in \mathbb{N}$ and $s = s(M) \in \mathbb{N}$ with the following property: For each $\varepsilon \in (0, 1/2)$, there is a neural network \tilde{x} with the following properties:

- \tilde{x} has at most $c \cdot \varepsilon^{-\theta}$ nonzero, (s, ε) -quantized weights;
- \tilde{x} has $2L + 8$ layers;
- $\forall x, y \in [-M, M]$, we have $|xy - R_\varrho(\tilde{x})(x, y)| \leq \varepsilon$
- $\forall x, y \in [-M, M]$ with $x \cdot y = 0$, we have $R_\varrho(\tilde{x})(x, y) = 0$

Approximation of Monomials (Lemma A.4)

Let $n, d, \ell \in \mathbb{N}$ be arbitrary. Then, there are constants $s = s(n) \in \mathbb{N}$, $c = c(d, n, \ell) \in \mathbb{N}$ and $L = L(d, n, \ell) \in \mathbb{N}$ such that $L \leq (1 + \lceil \log_2 n \rceil) \cdot (10 + \ell/d)$ with the following property: For each $\varepsilon \in (0, 1/2)$ and $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq n$, there is a neural network Φ_ε^α with d -dimensional input and one-dimensional output, with at most L layers, and with at most $c \cdot \varepsilon^{-d/\ell}$ nonzero, (s, ε) -quantized weights, and such that Φ_ε^α satisfies

$$|\mathbf{R}_\ell(\Phi_\varepsilon^\alpha)(x) - x^\alpha| \leq \varepsilon$$

for all $x \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d$

Approximation of Polynomials (Lemma A.5)

Let $d, m \in \mathbb{N}$, let $B, \beta > 0$,

let $\{c_{\ell, \alpha} : \ell \in \{1, \dots, m\}, \alpha \in \mathbb{N}_0^d, |\alpha| < \beta\} \subset [-B, B]$ be a sequence of coefficients, and let $(x_\ell)_{\ell=1}^m \subset [-1/2, 1/2]^d$ be a sequence of base points.

Then, there exist constants $c = c(d, \beta, B) > 0$, $s = s(d, \beta, B) \in \mathbb{N}$, and $L = L(d, \beta) \in \mathbb{N}$ with $L \leq 1 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \beta/d)$ such that $\forall \varepsilon \in (0, 1/2)$ there is a neural network Φ_ε^p with at most $c \cdot (\varepsilon^{-d/\beta} + m)$ many nonzero, (s, ε) -quantized weights, at most L layers, and with an m -dimensional output such that

$$\left| [R_\varrho(\Phi_\varepsilon^p)]_\ell(x) - \sum_{|\alpha| < \beta} c_{\ell, \alpha} \cdot (x - x_\ell)^\alpha \right| < \varepsilon \quad \text{for all } \ell \in \{1, \dots, m\} \text{ and } x \in [-1/2, 1/2]^d$$

Approximation of "cutoff" of a Network (Lemma A.6)

Let $d \in \mathbb{N}$, $p \in (0, \infty)$, and $B \geq 1$. Let $-1/2 \leq a_i \leq b_i \leq 1/2$ for $i = 1, \dots, d$, and let $\varepsilon \in (0, 1/2)$ be arbitrary. Then there exist constants $c = c(d) \in \mathbb{N}$, $s = s(d, B, p) \in \mathbb{N}$, and a neural network Λ_ε with a $d + 1$ -dimensional input, at most four layers, and at most c nonzero, (s, ε) -quantized weights such that for each neural network Φ with one-dimensional output layer and d -dimensional input layer, and with $\|R_\varrho(\Phi)\|_{L^\infty([-1/2, 1/2]^d)} \leq B$, we have

$$\left\| R_\varrho(\Lambda_\varepsilon)(\bullet, R_\varrho(\Phi)(\bullet)) - \chi_{\prod_{i=1}^d [a_i, b_i]} \cdot R_\varrho(\Phi) \right\|_{L^p([-1/2, 1/2]^d)} \leq \varepsilon$$

Approximation of "cutoff" of a Network (Lemma A.7)

Let $d, m, s \in \mathbb{N}$, $p \in (0, \infty)$, and $\varepsilon \in (0, 1/2)$, and let Φ be a neural network with d dimensional input and m -dimensional output, and with (s, ε) -quantized weights. Furthermore, let $B \geq 1$ with $\left\| \left[R_{\underline{e}}(\Phi) \right]_{\ell} \right\|_{L^{\infty}([-1/2, 1/2]^d)} \leq B$ for all $\ell = 1, \dots, m$. Finally, let $-1/2 \leq a_{i,\ell} \leq b_{i,\ell} \leq 1/2$ for $i = 1, \dots, d$ and $\ell = 1, \dots, m$. Then, there exist constants $c = c(d) > 0$, $s_0 = s_0(d, B, p) \in \mathbb{N}$, and a neural network Ψ_{ε} with d -dimensional input layer and 1-dimensional output layer, with at most $6 + L(\Phi)$ layers, and at most $c \cdot (m + L(\Phi) + M(\Phi))$ nonzero, $(\max\{s, s_0\}, \varepsilon/m)$ -quantized weights, such that

$$\left\| R_{\underline{e}}(\Psi_{\varepsilon}) - \sum_{\ell=1}^m \chi_{\Pi_{i=1}^d[a_{i,\ell}, b_{i,\ell}]} \cdot [R_{\underline{e}}(\Phi)]_{\ell} \right\|_{L^p([-1/2, 1/2]^d)} \leq \varepsilon$$

Taylor Approximation of Smooth Functions (Lemma A.8)

Let $\beta \in (0, \infty)$, and write $\beta = n + \sigma$ with $n \in \mathbb{N}_0$ and $\sigma \in (0, 1]$, and let $d \in \mathbb{N}$. Then there is a constant $C = C(\beta, d) > 0$ with the following property:

For each $f \in \mathcal{F}_{\beta, d, B}$ and arbitrary $x_0 \in (-1/2, 1/2)^d$, there is a polynomial $p(x) = \sum_{|\alpha| \leq n} c_\alpha (x - x_0)^\alpha$ with $c_\alpha \in [-C \cdot B, C \cdot B]$ for all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq n$ and such that

$$|f(x) - p(x)| \leq C \cdot B \cdot |x - x_0|^\beta \quad \text{for all } x \in [-1/2, 1/2]^d$$

In fact, $p = p_{f, x_0}$ is the Taylor polynomial of f of degree n

Proof of Theorem 3.1

Set $p_0 := \lceil p \rceil \in \mathbb{N}$. It suffices to show the result for approximation in L^{p_0} -norm since $\|f\|_{L^p([-1/2, 1/2]^d)} \leq \|f\|_{L^q([-1/2, 1/2]^d)}$ for $0 < p \leq q < \infty$.

Also, let $C = C(d, \beta) > 0$ be the constant from Lemma A.8 and define

$$N := \left\lceil \left(\frac{\varepsilon}{4CBd^\beta} \right)^{-\frac{1}{\beta}} \right\rceil \in \mathbb{N}.$$

Finally, for $\lambda \in \{1, \dots, N\}^d$, set

$$I_\lambda := \prod_{i=1}^d \left[\frac{\lambda_i - 1}{N} - \frac{1}{2}, \frac{\lambda_i}{N} - \frac{1}{2} \right]$$

Then, we have

$$\left[-\frac{1}{2}, \frac{1}{2} \right]^d = \bigcup_{\lambda \in \{1, \dots, N\}^d} I_\lambda \text{ and } I_\lambda \subset \bar{B}_{1/N}^{\|\cdot\|_{\ell^\infty}}(x) \subset \bar{B}_{d/N}^{\|\cdot\|_{\ell^\infty}}(x) \quad \forall x \in I_\lambda$$

Proof of Theorem 3.1 (Continued 1)

Also, write $\{1, \dots, N\}^d = \{\lambda_1, \dots, \lambda_{N^d}\}$, and for each $i \in \{1, \dots, N^d\}$, choose $x_i \in \text{interior}(I_{\lambda_i})$ and set $c_{i,\alpha} := \partial^\alpha f(x_i)/\alpha!$ for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq n$. Since $f \in \mathcal{F}_{\beta,d,B}$, $|c_{i,\alpha}| \leq B$.

By Lemma A.8, we have

$$\sup_{i \in \{1, \dots, N^d\}} |f(x) - p_{i,\alpha}(x)| \leq CB \left(\frac{d}{N} \right)^\beta$$

where $p_{i,\alpha}(x) := \sum_{|\alpha| \leq n} \frac{\partial^\alpha f(x_i)}{\alpha!} (x - x_i)^\alpha$, Taylor polynomial of f of degree n .

In particular,

$$|p_{i,\alpha}(x)| \leq |f(x)| + Cd^\beta B \leq \lceil (1 + Cd^\beta)B \rceil =: B_1 \quad \forall x \in I_{\lambda_i}$$

Proof of Theorem 3.1 (Continued 2)

For the polynomial defined above for each base point $(x_i)_{i=1,\dots,N^d}$, we can take $\Phi_{\varepsilon/4}^P$ by applying Lemma A.5 with $\varepsilon/4$ instead of *epsilon* and with $m = N^d$. This network has at most $L_1 = L_1(d, \beta) \leq 1 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \beta/d)$ layers and at most $c_1 (\varepsilon^{-d/\beta} + N^d)$ nonzero (s_1, ε) -quantized weights for some $c_1 = c_1(d, \beta, B) > 0$ and $s_1 = s_1(d, \beta, B) \in \mathbb{N}$.

Now apply Lemma A.1 with B_1 instead of B to obtain a network $\Psi_{\varepsilon/4}^P$ such that

$$R_\varrho \left(\Psi_{\varepsilon/4}^P \right) = (\tau_{B_1} \times \cdots \times \tau_{B_1}) \circ R_\varrho \left(\Phi_{\varepsilon/4}^P \right)$$

where $\tau_{B_1} : \mathbb{R} \rightarrow [-B_1, B_1]$ is 1-Lipschitz and satisfies $\tau_{B_1}(x) = x$ for all $x \in [-B_1, B_1]$ and has at most $L_2 := L(\Psi_{\varepsilon/4}^P) \leq 2 + L_1 \leq 3 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \beta/d)$ and at most $2c_1 \cdot (\varepsilon^{-d/\beta} + N^d) + c_2 \cdot N^d \leq c_3 \cdot (\varepsilon^{-d/\beta} + N^d)$ nonzero, (s_2, ε) -quantized weights for an absolute constant $c_2 > 0$ and suitable $s_2 = s_2(d, \beta, B) \in \mathbb{N}$ and $c_3 = c_3(d, \beta, B) > 0$.

Proof of Theorem 3.1 (Continued 3)

With these preparations, we have

$$\begin{aligned} \left| \left[R_{\varrho} \left(\Psi_{\varepsilon/4}^p \right) (x) \right]_i - p_{i,\alpha}(x) \right| &= \left| \tau_{B_1} \left(\left[R_{\varrho} \left(\Phi_{\varepsilon/4}^p \right) \right]_i (x) \right) - \tau_{B_1} (p_{i,\alpha}(x)) \right| \\ &\leq \left| \left[R_{\varrho} \left(\Phi_{\varepsilon/4}^p \right) \right]_i (x) - p_{i,\alpha}(x) \right| \leq \frac{\varepsilon}{4} \end{aligned}$$

for all $x \in I_{\lambda_i}$ and $i \in \{1, \dots, N^d\}$. Hence, we can write

$$\begin{aligned} \left\| f - \sum_{i \in \{1, \dots, N^d\}} \chi_{I_{\lambda_i}} \left[R_{\varrho} \left(\Psi_{\varepsilon/4}^p \right) \right]_i \right\|_{L^\infty} &\leq \sup_{i \in \{1, \dots, N^d\}} \left| f(x) - \left[R_{\varrho} \left(\Psi_{\varepsilon/4}^p \right) \right]_i (x) \right| \\ &\leq \frac{\varepsilon}{4} + \sup_{x \in I_{\lambda_i}} |f(x) - p_{i,\alpha}(x)| \\ &\leq \frac{\varepsilon}{4} + CB \left(\frac{d}{N} \right)^\beta \\ &\leq \frac{\varepsilon}{2} \end{aligned}$$

Proof of Theorem 3.1 (Continued 4)

Now, apply Lemma A.7 with $\varepsilon/2$ instead of ε , p_0 instead of p , $\Phi = \Psi_{\varepsilon/4}^p$, $m = N^d$, and $\prod_{i=1}^d [a_{i,\ell}, b_{i,\ell}] = I_{\lambda_\ell}$ to obtain a network Ψ_ε such that

$$\left\| R_\varrho(\Psi_\varepsilon) - \sum_{i \in \{1, \dots, N^d\}} \chi_{I_{\lambda_i}} \left[R_\varrho \left(\Psi_{\varepsilon/4}^p \right) \right]_i \right\|_{L^{p_0}} \leq \frac{\varepsilon}{2}$$

Then Ψ_ε has at most $3 + L_2 \leq 9 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \beta/d)$ layers and at most

$$\begin{aligned} c_4 \cdot \left(N^d + L \left(\Psi_{\varepsilon/4}^p \right) + M \left(\Psi_{\varepsilon/4}^p \right) \right) &\leq c_4 \cdot \left(N^d + L_2 + M \left(\Psi_{\varepsilon/4}^p \right) \right) \\ &\leq c_5 \cdot \left(N^d + c_3 \left(\varepsilon^{-d/\beta} + N^d \right) \right) \end{aligned}$$

nonzero, $(\max \{s_2, s_0\}, \varepsilon / (2N^d))$ -quantized weights, with constants $c_4 = c_4(d) > 0$, $c_5 = c_5(d, \beta) > 0$ and $s_0 = s_0(d, B, p)$.

By using the inequality $\varepsilon / (2N^d) \geq c_7 \cdot \varepsilon^{1+d/\beta}$ for some $c_7 = c_7(d, \beta, B)$, we can conclude that the weights of Ψ_ε is (s, ε) -quantized for $s = s(d, \beta, p, B)$

Proof of Theorem 3.1 (Continued 5)

Note that Ψ_ε now satisfies

$$\begin{aligned} \|f - R_\varrho(\Psi_\varepsilon)\|_{L^{p_0}([-\frac{1}{2}, \frac{1}{2}]^d)} &\leq \left\| f - \sum_{i \in \{1, \dots, N^d\}} \chi_{I_{\lambda_i}} \left[R_\varrho \left(\Psi_{\varepsilon/4}^p \right) \right]_i \right\|_{L^\infty} \\ &\quad + \left\| R_\varrho(\Psi_\varepsilon) - \sum_{i \in \{1, \dots, N^d\}} \chi_{I_{\lambda_i}} \left[R_\varrho \left(\Psi_{\varepsilon/4}^p \right) \right]_i \right\|_{L^{p_0}([-\frac{1}{2}, \frac{1}{2}]^d)} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

Lastly, we apply Lemma A.1 to Ψ_ε to obtain Φ_ε^f so that

$$R_\varrho(\Phi_\varepsilon^f) = \tau_B \circ R_\varrho(\Psi_\varepsilon)$$

Since, $\|f\|_{\text{sup}} \leq B$, we have $f = \tau_B \circ f$ so that

$$\left\| R_\varrho(\Phi_\varepsilon^f) - f \right\|_{L^{p_0}([-\frac{1}{2}, \frac{1}{2}]^d)} \leq \|R_\varrho(\Psi_\varepsilon) - f\|_{L^{p_0}([-\frac{1}{2}, \frac{1}{2}]^d)} \leq \varepsilon$$

It is easy to see that Φ_ε^f satisfies all the properties stated in the theorem.

Approximation of Horizon Functions

(Lemma 3.4, P., Voigtlaender 2018)

For any $p, \beta, B > 0$ and $d \in \mathbb{N}_{\geq 2}$ there exist constants $c = c(d, \beta, B, p) > 0$, and $s = s(d, \beta, B, p) \in \mathbb{N}$, such that for every function $f \in \mathcal{HF}_{\beta, d, B}$ and every $\varepsilon \in (0, 1/2)$ there is a neural network Φ_ε^f with at most $(2 + \lceil \log_2 \beta \rceil) \cdot (14 + 2\beta/d)$ layers, and at most $c \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero, (s, ε) -quantized weights, such that $\|R_\varrho(\Phi_\varepsilon^f) - f\|_{L^p([-1/2, 1/2]^d)} < \varepsilon$. Moreover, $0 \leq R_\varrho(\Phi_\varepsilon^f)(x) \leq 1$ for all $x \in \mathbb{R}^d$

Approximation of Heaviside function (Lemma A.2)

Let $d \in \mathbb{N}_{\geq 2}$ and $H := \chi_{[0, \infty) \times \mathbb{R}^{d-1}}$. For every $\varepsilon > 0$ there exists a neural network Φ_ε^H with two layers and five (nonzero) weights which only take values in $\{\varepsilon^{-1}, 1, -1\}$, such that $0 \leq R_\varrho(\Phi_\varepsilon^H) \leq 1$ and $|H(x) - R_\varrho(\Phi_\varepsilon^H)(x)| \leq \chi_{[0, \varepsilon] \times \mathbb{R}^{d-1}}(x)$ for all $x \in \mathbb{R}^d$. Moreover, $\|H - R_\varrho(\Phi_\varepsilon^H)\|_{L^p([-1/2, 1/2]^d)} \leq \varepsilon^{1/p}$ for all $p \in (0, \infty)$.

Proof of Lemma 3.4

Write $f = H \circ \tilde{\gamma}$ where

$$\tilde{\gamma}(x) = (x_1 + \gamma(x_2, \dots, x_d), x_2, \dots, x_d) \text{ for } x = (x_1, \dots, x_d) \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d$$

where $\gamma \in \mathcal{F}_{\beta, d-1, B}$.

Apply Theorem 3.1 to approximate γ with $p = 1$, $d - 1$ instead of d , and $\frac{1}{2} \cdot \left(\frac{\varepsilon}{4}\right)^p$ instead of ε . This yields a network Φ_ε^γ with at most

$$L = L(d, \beta) \leq 11 + (1 + \lceil \log_2 \beta \rceil) \cdot \left(11 + \frac{\beta}{d-1}\right) \leq 11 + (1 + \lceil \log_2 \beta \rceil) \cdot \left(11 + \frac{2\beta}{d}\right)$$

layers and at most $c \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero, (s, ε) -quantized weights for $c = c(d, \beta, B, p) > 0$ and $s = s(d, \beta, B, p)$ such that

$$\|\mathbf{R}_\varrho(\Phi_\varepsilon^\gamma) - \gamma\|_{L^1} \leq \frac{1}{2} \cdot \left(\frac{\varepsilon}{4}\right)^p$$

Then, denote $\gamma_\varepsilon := \mathbf{R}_\varrho(\Phi_\varepsilon^\gamma)$

Proof of Lemma 3.4 (Continued 1)

Clearly, one can construct $\Phi_{\varepsilon}^{\tilde{\gamma}}$ with same number of nonzero weights and quantization (up to multiplicative constant dependent on d, β) and $L(\Phi_{\varepsilon}^{\tilde{\gamma}}) \leq 1 + L(\Phi_{\varepsilon}^{\gamma})$ such that

$$R_{\varrho} \left(\Phi_{\varepsilon}^{\tilde{\gamma}} \right) (x) = (x_1 + \gamma_{\varepsilon}(x_2, \dots, x_d), x_2, \dots, x_d) \quad \forall x \in \mathbb{R}^d$$

As a second step, apply Lemma A.2 with $\varepsilon' \in 2^{-\mathbb{N}}$ such that $\frac{1}{4} \cdot \left(\frac{\varepsilon}{4}\right)^p \leq \varepsilon' \leq \frac{1}{2} \cdot \left(\frac{\varepsilon}{4}\right)^p$ to obtain $\Phi_{\varepsilon'}^H$ with two layers and 5 nonzero weights such that

$$|H(x) - R_{\varrho} \left(\Phi_{\varepsilon'}^H \right)| \leq \chi_{[0, \varepsilon'] \times \mathbb{R}^{d-1}}(x)$$

and $0 \leq R_{\varrho} \left(\Phi_{\varepsilon'}^H \right) \leq 1$. We also have that all weights of $\Phi_{\varepsilon'}^H$ belong to $[-4(\varepsilon/4)^{-p}, 4(\varepsilon/4)^{-p}] \cap \mathbb{Z} \subset [-\varepsilon^{-s'}, \varepsilon^{-s'}] \cap \mathbb{Z}$ where $s' := 2 + 3[p]$.

Proof of Lemma 3.4 (Continued 2)

Now we concatenate two networks above to get $\Phi_{\varepsilon'}^H \odot \Phi_{\varepsilon}^{\tilde{\gamma}}$ with at most $\tilde{L} \leq 2 + L(\Phi_{\varepsilon}^{\tilde{\gamma}}) \leq 14 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \frac{2\beta}{d})$ layers and no more than $\tilde{c} \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero (s'', ε) -quantized weights for suitable $s'' = s''(d, \beta, B, p) \in \mathbb{N}$. Clearly, $0 \leq R_{\varrho}(\Phi_{\varepsilon'}^H \odot \Phi_{\varepsilon}^{\tilde{\gamma}}) \leq 1$.

Now, we check that $R_{\varrho}(\Phi_{\varepsilon'}^H \odot \Phi_{\varepsilon}^{\tilde{\gamma}})$ approximates $f = H \circ \tilde{\gamma}$ with L^p -error at most ε .

Now, observe that for $q := 1 + \frac{1}{p} \geq \max\{1, \frac{1}{p}\}$,

$$\begin{aligned} \left\| H \circ \tilde{\gamma} - R_{\varrho}(\Phi_{\varepsilon'}^H \odot \Phi_{\varepsilon}^{\tilde{\gamma}}) \right\|_{L^p} &= \left\| H \circ \tilde{\gamma} - R_{\varrho}(\Phi_{\varepsilon'}^H) \circ R_{\varrho}(\Phi_{\varepsilon}^{\tilde{\gamma}}) \right\|_{L^p} \\ &\leq 2^q \cdot \max \left\{ \left\| H \circ \tilde{\gamma} - H \circ R_{\varrho}(\Phi_{\varepsilon'}^H) \right\|_{L^p}, \left\| H \circ R_{\varrho}(\Phi_{\varepsilon'}^H) - R_{\varrho}(\Phi_{\varepsilon'}^H \odot \Phi_{\varepsilon}^{\tilde{\gamma}}) \right\|_{L^p} \right\} \\ &=: 2^q \cdot \max\{I, II\} \end{aligned}$$

From now, write $\chi_{\tilde{\gamma}}$ to denote indicator function of the set $\{x \in [-1/2, 1/2]^d : \tilde{\gamma}(x) > 0\}$, and we write $R_{\varrho}(\Phi_{\varepsilon}^{\tilde{\gamma}})_1$ to denote the first coordinate of the value $R_{\varrho}(\Phi_{\varepsilon}^{\tilde{\gamma}})$.

Proof of Lemma 3.4 (Continued 3)

We first bound I from above.

$$\begin{aligned}
 (2^q \cdot I)^p &= 2^{1+p} \cdot \left\| H \circ \tilde{\gamma} - H \circ R_\varrho(\Phi_{\varepsilon'}^H) \right\|_{L^p}^p \\
 &= 2^{1+p} \int_{[-1/2, 1/2]^d} \left| \chi_{\tilde{\gamma}_1 \geq 0}(x) - \chi_{R_\varrho(\Phi_{\varepsilon'}^H)_1 \geq 0}(x) \right|^p dx \\
 &= 2^{1+p} \int_{[-\frac{1}{2}, \frac{1}{2}]^{d-1}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{\tilde{\gamma}_1 \geq 0, R_\varrho(\Phi_{\varepsilon'}^H)_1 < 0}(x_1, \dots, x_d) \\
 &\quad + \chi_{\tilde{\gamma}_1 < 0, R_\varrho(\Phi_{\varepsilon'}^H)_1 \geq 0}(x_1, \dots, x_d) dx_1 d(x_2, \dots, x_d)
 \end{aligned}$$

Then, since

$$\begin{aligned}
 \chi_{\tilde{\gamma}_1 \geq 0, R_\varrho(\Phi_{\varepsilon'}^H)_1 < 0} = 1 &\iff x_1 + \gamma(x_2, \dots, x_d) \geq 0 \text{ and } x_1 + \gamma_\varepsilon(x_2, \dots, x_d) < 0 \\
 &\iff x_1 \in [-\gamma(x_2, \dots, x_d), -\gamma_\varepsilon(x_2, \dots, x_d))
 \end{aligned}$$

we have

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{\tilde{\gamma}_1 \geq 0, R_\varrho(\Phi_{\varepsilon'}^H)_1 < 0}(x_1, \dots, x_d) dx_1 \leq \max\{0, \gamma(x_2, \dots, x_d) - \gamma_\varepsilon(x_2, \dots, x_d)\}$$

Proof of Lemma 3.4 (Continued 4)

By similar reasoning, we also have

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{\tilde{\gamma}_1 < 0, R_\ell(\Phi_{\varepsilon'}^H)_1 \geq 0}(x_1, \dots, x_d) dx_1 \leq \max\{0, \gamma_\varepsilon(x_2, \dots, x_d) - \gamma(x_2, \dots, x_d)\}$$

Hence,

$$\begin{aligned} (2^q \cdot I)^p &\leq 2^{1+p} \int_{[-\frac{1}{2}, \frac{1}{2}]^{d-1}} \max\{0, \gamma(y) - \gamma_\varepsilon(y)\} + \max\{0, \gamma_\varepsilon - \gamma(y)\} dy \\ &= 2^{1+p} \|\gamma - \gamma_\varepsilon\|_{L^1([- \frac{1}{2}, \frac{1}{2}]^{d-1})} \\ &\leq \left(\frac{\varepsilon}{2}\right)^p \end{aligned}$$

This shows $2^q \cdot I \leq \frac{\varepsilon}{2}$.

Now we bound II . By construction of $R_\ell(\Phi_{\varepsilon'}^H)$, we have $|H(x) - R_\ell(\Phi_{\varepsilon'}^H)(x)| \leq \chi_{[0, \varepsilon'] \times \mathbb{R}^{d-1}}(x) \leq \chi_{[0, \frac{1}{2}(\frac{\varepsilon}{4})] \times \mathbb{R}^{d-1}}(x) \quad \forall x \in \mathbb{R}^d$.

Proof of Lemma 3.4 (Continued 5)

Therefore, we can write

$$\begin{aligned}
 (2^q \cdot II)^p &\leq 2^{1+p} \int_{[-\frac{1}{2}, \frac{1}{2}]^d} \chi_{0 \leq R_\varrho(\Phi_\varepsilon^{\tilde{\gamma}})_1 \leq \frac{1}{2}(\frac{\varepsilon}{4})^p} \times \mathbb{R}^{d-1}(x) \\
 &= 2^{1+p} \int_{[-\frac{1}{2}, \frac{1}{2}]^{d-1}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{0 \leq x_1 + \gamma_\varepsilon(x_2, \dots, x_d) \leq \frac{1}{2}(\varepsilon/4)^p} dx_1 d(x_2, \dots, x_d) \\
 &\leq 2^p \int_{[-\frac{1}{2}, \frac{1}{2}]^{d-1}} (\varepsilon/4)^p d(x_2, \dots, x_d) = (\varepsilon/2)^p
 \end{aligned}$$

from which we can conclude $2^q \cdot II \leq \varepsilon/2$.

Hence,

$$\left\| H \circ \tilde{\gamma} - R_\varrho(\Phi_\varepsilon^H \odot \Phi_\varepsilon^{\tilde{\gamma}}) \right\|_{L^p} \leq 2^q \cdot \max\{I, II\} \leq \varepsilon/2 < \varepsilon$$

Finally, noting that we can change coordinates by applying a permutation matrix to A_1 for any neural network $\Phi = ((A_1, b_1), \dots, (A_L, b_L))$ without changing number of layers, number of nonzero weights, and quantization, we see that we can assume T in the definition of $\mathcal{HF}_{\beta, d, B}$ to be the identity matrix without loss of generality, so we're done. \square

Approximation of Piecewise Constant Functions (Theorem 3.5 P., Voigtlaender 2018)

For $r \in \mathbb{N}$, $d \in \mathbb{N}_{\geq 2}$, and $p, \beta, B > 0$, there are constants $c = c(d, r, p, \beta, B) > 0$ and $s = s(d, r, p, \beta, B) \in \mathbb{N}$, such that for any $K \in \mathcal{K}_{r, \beta, d, B}$ and any $\varepsilon \in (0, 1/2)$, there is a neural network Φ_ε^K with at most $(3 + \lceil \log_2 \beta \rceil) \cdot (11 + 2\beta/d)$ layers, and at most $c \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero, (s, ε) -quantized weights such that

$$\left\| R_\varrho \left(\Phi_\varepsilon^K \right) - \chi_K \right\|_{L^p([-1/2, 1/2]^d)} < \varepsilon \quad \text{and} \quad \left\| R_\varrho \left(\Phi_\varepsilon^K \right) \right\|_{\text{sup}} \leq 1$$

Proof of Theorem 3.5

For $\lambda = (\lambda_1, \dots, \lambda_d) \in \{1, \dots, 2^r\}^d$, define

$$I_\lambda := \prod_{i=1}^d \left[(\lambda_i - 1) \cdot 2^{-r} - \frac{1}{2}, \lambda_i \cdot 2^{-r} - \frac{1}{2} \right]$$

We have by construction (with disjointness up to null sets) that $[-1/2, 1/2]^d =$

$$\bigcup_{\lambda \in \{1, \dots, 2r\}^d} I_\lambda, \quad \text{and} \quad I_\lambda \subset \bar{B}_{2^{-r}}^{\|\cdot\|_{e^\infty}}(x) \text{ for all } x \in I_\lambda.$$

Then, by definition of $\mathcal{K}_{r,\beta,d,B}$, for each $\lambda \in \{1, \dots, 2^r\}^d$, $\exists f_\lambda \in \mathcal{HF}_{\beta,d,B}$ such that $\chi_{I_\lambda} \chi_K = \chi_{I_\lambda} f_\lambda$

Applying Lemma 3.4 to each f_λ , we get Φ_ε^λ such that

$$\left\| R_\varrho \left(\Phi_\varepsilon^\lambda \right) - f_\lambda \right\|_{L^p} \leq \frac{\varepsilon}{2^{1+q+rdq}} \quad \text{and} \quad 0 \leq R_\varrho \left(\Phi_\varepsilon^\lambda \right) (x) \leq 1 \quad \forall x \in \mathbb{R}^d$$

and Φ_ε^λ has at most $L_1 \leq 14 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \frac{2\beta}{d})$ layers and at most $c_1 \cdot \varepsilon^{p(d-1)/\beta}$ nonzero, (s_1, ε) -quantized weights for $L_1 = L_1(d, \beta)$, $c_1 = c_1(d, \beta, B, r, p) > 0$, $s_1 = s_1(d, \beta, B, r, p) \in \mathbb{N}$

Proof of Theorem 3.5 (Continued 1)

There is no problem in assuming Φ_ε^λ has exactly L_1 layers by using sparse concatenation with identity network, and we'd still have same order of complexity and number of layers.

Then, we can parallelize all of Φ_ε^λ to obtain

$$\Phi := P(\Phi_\varepsilon^{\lambda_1}, P(\Phi_\varepsilon^{\lambda_2}, \dots, P(\Phi_\varepsilon^{\lambda_{2^{rd}-1}}, \Phi_\varepsilon^{\lambda_{2^{rd}}}) \dots))$$

Apply Lemma A.7 with $m = 2^{rd}$, $B = 1$ with $\varepsilon/(2^{1+q})$ instead of ε and with I_{λ_l} for $l \in \{1, \dots, 2^{rd}\}$ to obtain Ψ that satisfies

$$\begin{aligned} \|R_\varrho(\Psi) - \chi_K\|_{L^p} &= \left\| R_\varrho(\Psi) - \sum_{l \in \{1, \dots, 2^{rd}\}} \chi_{I_{\lambda_l}} f_{\lambda_l} \right\|_{L^p} \\ &\leq 2^q \left\| R_\varrho(\Psi) - \sum_{l \in \{1, \dots, 2^r\}} \chi_{I_{\lambda_l}} [R_\varrho(\Phi)]_l \right\|_{L^p} + 2^q \left\| \sum_{l \in \{1, \dots, 2^{rd}\}} \chi_{I_{\lambda_l}} ([R_\varrho(\Phi)]_l - f_{\lambda_l}) \right\|_{L^p} \\ &\leq \frac{\varepsilon}{2} + 2^{q+rdq} \cdot \max_{l \in \{1, \dots, 2^{rd}\}} \left\{ \left\| R_\varrho(\Phi_\varepsilon^{\lambda_l}) - f_{\lambda_l} \right\|_{L^p} \right\} \leq \varepsilon \end{aligned}$$

Proof of Theorem 3.5 (Continued 2)

By properties of parallelization, we observe Φ has L_1 layers and at most $2^{rd} \cdot c_1 \cdot \varepsilon^{p(d-1)/\beta}$ nonzero (s_1, ε) -quantized weights.

By Lemma A.7, Ψ has at most $6+L(\Phi) = 6+L_1$ layers and $(\max\{s_0, s_1\}, \varepsilon/(2^{1-L_1}))$ -quantized weights so that we can say it is also (s_2, ε) -quantized for $s_2 = s_2(d, \beta, B, r, p) \in \mathbb{N}$. Finally, $M(\Psi) \leq c \cdot (2^{rd} + L_1 + M(\Phi)) \leq c_3 \cdot \varepsilon^{-p(d-1)/\beta}$ for $c_3 = c_3(d, \beta, B, r, p) > 0$.

Finally, we apply Lemma A.1 to bound the output of the network by 1, which surely doesn't affect the order of complexity and number of layers upto multiplicative constants. \square

Approximation of Piecewise Smooth Functions (Corollary 3.7, P., Voigtlaender 2018)

Let $r \in \mathbb{N}$, $d \in \mathbb{N}_{\geq 2}$, and $p, B, \beta > 0$. Let $\beta' := \frac{d\beta}{p(d-1)}$, and set $\beta_0 := \max\{\beta, \beta'\}$. Then there exist constants $c = c(d, p, \beta, r, B) > 0$ and $s = s(d, p, \beta, r, B) \in \mathbb{N}$, such that for all $\varepsilon \in (0, 1/2)$ and all $f \in \mathcal{E}_{r, \beta, d, B}^p$ there is a neural network Φ_ε^f with at most $(4 + \lceil \log_2 \beta_0 \rceil) \cdot (12 + 3\beta_0/d)$ layers, and at most $c \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero, (s, ε) -quantized weights, such that

$$\left\| \mathbf{R}_{\varrho} \left(\Phi_{\varepsilon}^f \right) - f \right\|_{L^p([-1/2, 1/2]^d)} \leq \varepsilon \quad \text{and} \quad \left\| \mathbf{R}_{\varrho} \left(\Phi_{\varepsilon}^f \right) \right\|_{\text{sup}} \leq \lceil B \rceil$$

Proof of Corollary 3.7

Let $q := \max\{1, p^{-1}\}$, $\varepsilon \in (0, 1/2)$ and $f = \chi_K \cdot g$ with $g \in \mathcal{F}_{\beta', d, B}$ and $K \in \mathcal{K}_{r, \beta, d, B}$. We prove this corollary in 4 steps.

First, apply Theorem 3.5 with $\frac{\varepsilon}{3 \cdot 4^q \cdot B}$ instead of ε to obtain a neural network Φ_ε^K with at most

$$L_1 \leq 22 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \frac{2\beta}{d}) \leq 22 + (1 + \lceil \log_2 \beta_0 \rceil) \cdot (11 + \frac{2\beta_0}{d})$$

layers and at most $c_1 \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero (s_1, ε) -quantized weights for $c_1 = c_1(d, \beta, r, p, B)$, $s_1 = s_1(d, \beta, r, p, B)$, and $L_1 = L_1(d, \beta)$ such that

$$\left\| R_\varrho \left(\Phi_\varepsilon^K \right) - \chi_K \right\|_{L^p([-1/2, 1/2]^d)} \leq \frac{\varepsilon}{3 \cdot 4^q \cdot B} \quad \text{and} \quad \left\| R_\varrho \left(\Phi_\varepsilon^K \right) \right\|_{\text{sup}} \leq 1$$

Proof of Corollary 3.7 (Continued 1)

Second, apply Theorem 3.1 with $\frac{\varepsilon}{3 \cdot 4^q}$ instead of ε to obtain a neural network Φ_{η}^g with at most

$$L_2 \leq 11 + (1 + \lceil \log_2 \beta' \rceil) \cdot (11 + \frac{\beta'}{d}) \leq 11 + (1 + \lceil \log_2 \beta_0 \rceil) \cdot (11 + \frac{2\beta_0}{d})$$

layers and at most $c_2 \cdot \varepsilon^{-d/\beta'} = c_2 \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero (s_2, ε) -quantized weights for $c_2 = c_2(d, \beta, r, p, B)$, $s_2 = s_2(d, \beta, r, p, B)$, and $L_2 = L_2(d, \beta, p)$ such that

$$\|R_\varrho(\Phi_\varepsilon^g) - g\|_{L^p([-1/2, 1/2]^d)} \leq \frac{\varepsilon}{3 \cdot 4^q} \quad \text{and} \quad \|R_\varrho(\Phi_\varepsilon^g)\|_{\text{sup}} \leq \lceil B \rceil$$

By the usual method of sparse concatenation of identity layers, we can assume

$$L(\phi_{\varepsilon}^K) = L\phi_{\varepsilon}^g = \max\{L_1, L_2\} \leq 22 + (1 + \lceil \log_2 \beta_0 \rceil) \cdot (11 + 2\beta_0/d)$$

with only possible change in the constants c_1 and/or c_2 .

Proof of Corollary 3.7 (Continued 3)

Finally, we check that Ψ^f satisfies the claimed error bound.

$$\begin{aligned}
& \left\| R_\varrho \left(\Psi_\varepsilon^f \right) - f \right\|_{L^p} = \left\| R_\varrho(\tilde{x}) \left(R_\varrho \left(\Phi_\varepsilon^K \right), R_\varrho \left(\Phi_\varepsilon^g \right) \right) - f \right\|_{L^p} \\
& \leq 2^q \cdot \left\| R_\varrho(\tilde{x}) \left(R_\varrho \left(\Phi_\varepsilon^K \right), R_\varrho \left(\Phi_\varepsilon^g \right) \right) - R_\varrho \left(\Phi_\varepsilon^K \right) \cdot R_\varrho \left(\Phi_\varepsilon^g \right) \right\|_{L^p} \\
& \quad + 2^q \cdot \left\| R_\varrho \left(\Phi_\varepsilon^K \right) \cdot R_\varrho \left(\Phi_\varepsilon^g \right) - f \right\|_{L^p} \\
& \leq \frac{\varepsilon}{3} + 4^q \cdot \left\| R_\varrho \left(\Phi_\varepsilon^K \right) \cdot [R_\varrho \left(\Phi_\varepsilon^g \right) - g] \right\|_{L^p} + 4^q \cdot \left\| g \cdot [R_g \left(\Phi_\varepsilon^K \right) - \chi_K] \right\|_{L^p} \\
& \leq \frac{\varepsilon}{3} + 4^q \cdot \| R_\varrho \left(\Phi_\varepsilon^g \right) - g \|_{L^p} + 4^q B \cdot \left\| R_g \left(\Phi_\varepsilon^K \right) - \chi_K \right\|_{L^p} \\
& \leq \varepsilon
\end{aligned}$$

where second to last inequality used that $\|\mathbf{R}_{\varrho}(\Phi_{\varepsilon}^K)\|_{\text{sup}} \leq 1$ and $\|g\|_{\text{sup}} \leq B$. \square