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

Ko, Suh, Huo

Georgia Tech

Summer of 2020

Table of Contents

Basics of Neural Networks

Definition of Neural Network

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

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

Then, a neural network is

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

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

$$\begin{aligned} x_0 &:= x \\ x_l &= \varrho(A_l x_{l-1} + b_l) \text{ for } l = 1, \ldots, L-1 \\ x_L &= A_L x_{l-1} + b_L \end{aligned}$$

Complexity of Neural Network

There are generally two ways of describing complexity of a neural network. Given the neural network Φ as above,

Number of nonzero weights, layers, and neurons

- $M(\Phi) := \sum_{j=1}^{L} (\|A_l\|_0 + \|b_l\|_0)$ (Number of nonzero weights)
- $L(\Phi) := L$ (Number of layers)
- $N(\Phi) := \sum_{k=0}^{L} N_k$ (Number of neurons)

Quantization

A neural network Φ is (s, ε)-quantized if all weights (elements of A_i, b_i for $i \in \{1, \ldots, L\}$) are elements of

$$[-\epsilon^{-s}, \epsilon^{-s}] \cap 2^{-s\lceil \log_2(1/\varepsilon) \rceil} \mathbb{Z}$$

.

Operations of Neural Network

Realization of Identity

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

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

where

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

Then, $\mathrm{R}_\varrho(\Phi_d^{\mathrm{Id}}) = \mathrm{Id}_{\mathbb{R}^d}$

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

Operations of Neural Networks

Concatenation

Let $L_1, L_2 \in \mathbb{N}$ and let $\Phi^1 = \left((A_1^1, b_1^1), \ldots, (A_{L_1}^1, b_{L_1}^1) \right)$ and $\Phi^2 = \left((A_1^2, b_1^2), \ldots, (A_{L_2}^2, b_{L_2}^2) \right)$ 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:

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

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

Sparse Concatenation

Let $\varrho : \mathbb{R} \to \mathbb{R}$ 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

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

Operations of Neural Networks

Properties of sparse concatenation

It follows immediately from above definitions that

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

Then, we can deduce the following properties:

- $L(\Phi^1 \odot \Phi^2) = L_1 + L_2$
- $\mathrm{R}_\varrho\left(\Phi^1 \odot \Phi^2\right) = \mathrm{R}_\varrho\left(\Phi^1\right) \circ \mathrm{R}_\varrho\left(\Phi^2\right)$
- $M\left(\Phi^1 \odot \Phi^2\right) \leq 2M\left(\Phi^1\right) + 2M\left(\Phi^2\right)$

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

$$M\left(\Phi^1 \odot \cdots \odot \Phi^n\right) \leq 4^{n-1} \cdot \max\left\{M\left(\Phi^1\right), \ldots, M\left(\Phi^n\right)\right\}$$

Operations of Neural Networks

Parallelization

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

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

where

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

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

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

Parallelization

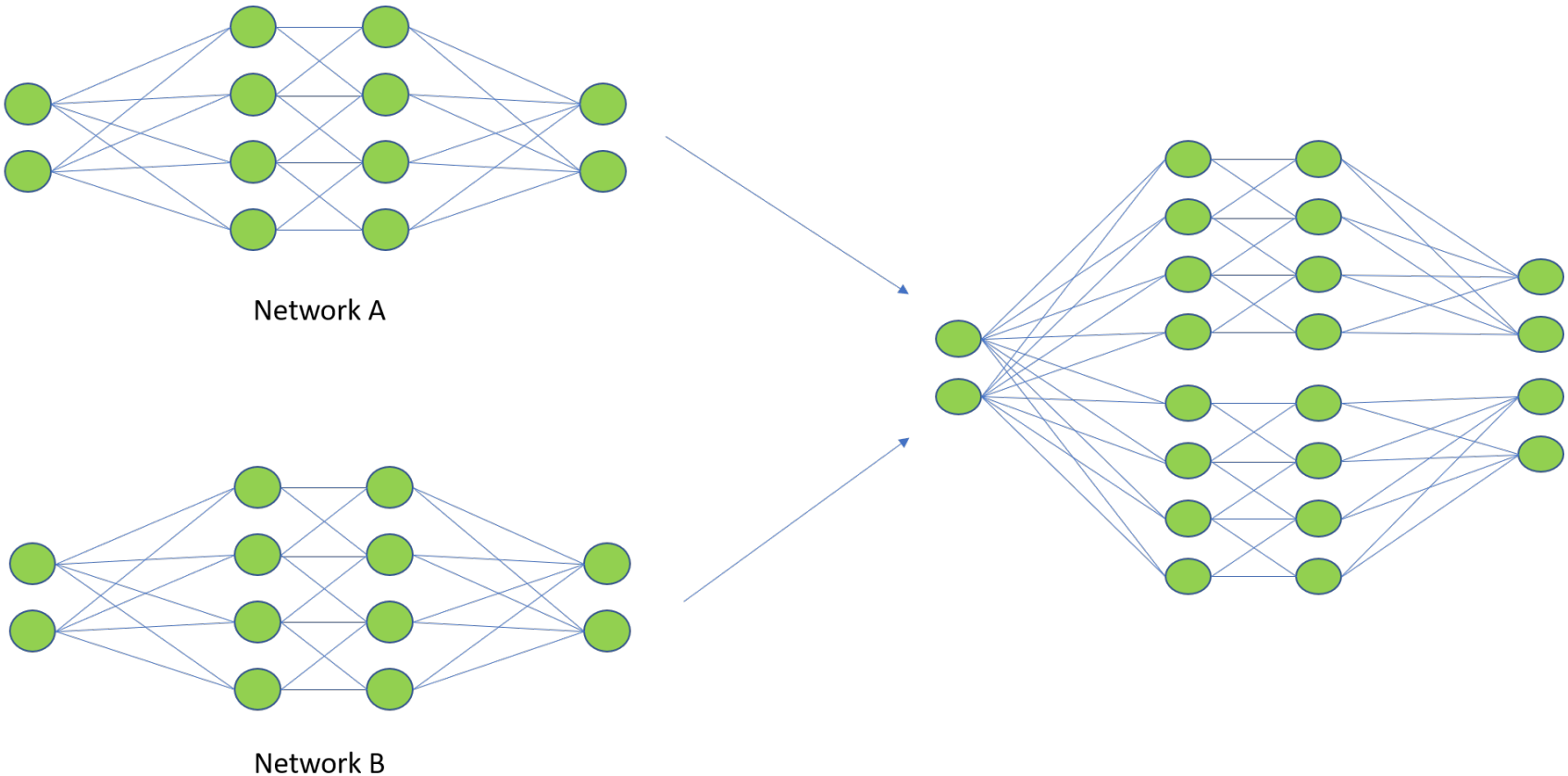

Figure 1: Parallelization.

Problem

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 $\mathrm{NNs}\left(\Phi_n^\varrho\right)_{n\in\mathbb{N}}$ such that $\|f - \Phi_n^\varrho\|_{L^\infty} \lesssim N\left(\Phi_n^\varrho\right)^{-1} \to 0$
- Mhaskar; 1993: For $\mathcal{C} = C^s\left([0,1]^d\right)$, ϱ a sigmoidal function of order $k \geq 2$, and $L\left(\Phi_n^\varrho\right) = L(k,d,s)$, there exist $\mathrm{NNs}\left(\Phi_n^\varrho\right)_{n\in\mathbb{N}}$ such that $\|f - \Phi_n^\varrho\|_{L^\infty} \lesssim N\left(\Phi_n^\varrho\right)^{-s/d} \to 0$
- Yarotsky; 2017: For $\mathcal{C} = C^s\left([0,1]^d\right)$ and $\varrho(x) = \max\{0,x\}$ (called ReLU) and $L\left(\Phi_n^\varrho\right) \asymp \log(n)$ we have that there exist $\mathrm{NNs}\left(\Phi_n^\varrho\right)_{n\in\mathbb{N}}$ such that $\|f - \Phi_n^\varrho\|_{L^\infty} \lesssim W\left(\Phi_n^\varrho\right)^{-s/d} \to 0$

Function Classes

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

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

where

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

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

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

Horizon Functions and Piecewise Constant Functions

Horizon functions

We define the class of horizon functions as:

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

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

Piecewise constant functions

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

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

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.

Piecewise Smooth Functions

Piecewise smooth functions

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

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

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 $\left[-\frac{1}{2}, \frac{1}{2}\right]^d$.

Theorem 1

Theorem 1: 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

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

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. (Lemma A.3, A.4, A.5)
- 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. (Lemma A.8)
- 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. (Lemma A.5)
- Construct a neural network with 1-dimensional output that approximates each output of the network from previous step restricted to the corresponding grid cubes. (Lemma A.6, A.7)
- 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. (Lemma A.1)

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

$$s_0 := \max\left\{\left\lceil \log_2(\lceil B \rceil)\right\rceil, s\right\}$$

- $\mathrm{R}_\varrho(\Phi) = (\tau_B \times \cdots \times \tau_B) \circ \mathrm{R}_\varrho(\Psi)$, where the function

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

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

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 $\widetilde{\times}$ with the following properties:

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

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

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

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 $\left\{c_{\ell,\alpha} : \ell \in \{1,\dots,m\}, \alpha \in \mathbb{N}_0^d, |\alpha| < \beta\right\} \subset [-B, B]$ 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 $\Phi_\varepsilon^{\mathrm{p}}$ with at most $c \cdot \left(\varepsilon^{-d/\beta} + m\right)$ many nonzero, (s, ε) -quantized weights, at most L layers, and with an m -dimensional output such that $\left|\left[\mathrm{R}_\varrho\left(\Phi_\varepsilon^{\mathrm{p}}\right)\right]_\ell(x) - \sum_{|\alpha|<\beta} c_{\ell,\alpha} \cdot (x - x_\ell)^\alpha\right| < \varepsilon$ for all $\ell \in \{1,\dots,m\}$ 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 $i = 1, \ldots, d$, 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 $\|\mathrm{R}_\varrho(\Phi)\|_{L^\infty([-1/2,1/2]^d)} \leq B$, we have

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

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 $\left\| \left[\mathrm{R}_{\underline{\varrho}}(\Phi) \right]_\ell \right\|_{L^\infty([-1/2,1/2]^d)} \leq B$ for all $\ell = 1, \ldots, m$. Finally, let $-1/2 \leq a_{i,\ell} \leq b_{i,\ell} \leq 1/2$ for $i = 1, \ldots, d$ and $\ell = 1, \ldots, m$. 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

$$\| \mathrm{R}_\varrho(\Psi_\varepsilon) - \sum_{\ell=1}^{m} \chi_{\prod_{i=1}^d [a_{i,\ell}, b_{i,\ell}]} \cdot [\mathrm{R}_\varrho(\Phi)]_\ell \|_{L^p([-1/2,1/2]^d)} \leq \varepsilon$$

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

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

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

[Proof of Theorem 3.1] Step 1: Gridifying

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 $\lambda \in \{1, \ldots, N\}^d$, 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

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

Step 2: Taylor Approximation of f

Also, write $\{1,\dots,N\}^d = \{\lambda_1,\dots,\lambda_{N^d}\}$, and for each $i \in \{1,\dots,N^d\}$, choose $x_i \in \mathit{interior}(I_{\lambda_i})$ and set $c_{i,\alpha} := \partial^\alpha f(x_i)/\alpha!$ for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le n$. Since $f \in \mathcal{F}_{\beta,d,B}$, $|c_{i,\alpha}| \le B$.
By Lemma A.8, we have

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

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

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

Step 3: Construction of NN (# outputs = # grids)

For the polynomial defined above for each base point $(x_i)_{i=1,\dots,N^d}$, we can take $\Phi^p_{\varepsilon/4}$ by applying Lemma A.5 with $\varepsilon/4$ instead of ϵ 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 $c_1 \left(\varepsilon^{-d/\beta} + N^d\right)$ 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 $\Psi^p_{\varepsilon/4}$ such that

$$\mathrm{R}_\varrho\left(\Psi^p_{\varepsilon/4}\right) = (\tau_{B_1} \times \cdots \times \tau_{B_1}) \circ \mathrm{R}_\varrho\left(\Phi^P_{\varepsilon/4}\right)$$

where $\tau_{B_1} : \mathbb{R} \to [-B_1, B_1]$ is 1-Lipschitz and satisfies $\tau_{B_1}(x) = x$ for all $x \in [-B_1, B_1]$ and has at most $L_2 := L(\Psi^p_{\varepsilon/4}) \leq 2 + L_1 \leq 3 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \beta/d)$ and at most $2c_1 \cdot \left(\varepsilon^{-d/\beta} + N^d\right) + c_2 \cdot N^d \leq c_3 \cdot \left(\varepsilon^{-d/\beta} + N^d\right)$ 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 $c_3 = c_3(d,\beta,B) > 0$.

Step 4: Approximation of f with NN

With these preparations, we have

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

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

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

Step 5: Refinement of NN using "cutoff" result

Now, apply Lemma A.7 with $\varepsilon/2$ instead of ε, p_0 instead of p, $\Phi = \Psi^p_{\varepsilon/4}$, $m = N^d$, and $\prod_{i=1}^d [a_{i,\ell}, b_{i,\ell}] = I_{\lambda_\ell}$ to obtain a network Ψ_ε such that $\left\| \mathrm{R}_\varrho(\Psi_\varepsilon) - \sum_{i\in\{1,\dots,N^d\}} \chi_{I_{\lambda_i}} \left[\mathrm{R}_\varrho\left(\Psi^{\mathrm{p}}_{\varepsilon/4}\right)\right]_i \right\|_{L^{p_0}} \leq \frac{\varepsilon}{2}$
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

$$c_4 \cdot \left(N^d + L\left(\Psi^{\mathrm{p}}_{\varepsilon/4}\right) + M\left(\Psi^{\mathrm{p}}_{\varepsilon/4}\right)\right) \leq c_4 \cdot \left(N^d + L_2 + M\left(\Psi^{\mathrm{p}}_{\varepsilon/4}\right)\right)$$
$$\leq c_5 \cdot \left(N^d + c_3\left(\varepsilon^{-d/\beta} + N^d\right)\right)$$

nonzero, $\left(\max\{s_2, s_0\}, \varepsilon/\left(2N^d\right)\right)$-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 $s = s(d,\beta,p,B)$

Step 6: Final approximation with bounded outputs

Note that Ψ_ε now satisfies

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

Lastly, we apply Lemma A.1 to Ψ_ε to obtain Φ^f_ε so that

$$\mathrm{R}_\varrho(\Phi^f_\varepsilon) = \tau_B \circ \mathrm{R}_\varrho(\Psi_\varepsilon)$$

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

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

It is easy to see that Φ^f_ε satisfies all the properties stated in the theorem.

Theorem 2

Theorem 2: 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 $\left\| \mathrm{R}_\varrho\left(\Phi_\varepsilon^f\right) - f \right\|_{L^p([-1/2,1/2]^d)} < \varepsilon$. Moreover, $0 \leq \mathrm{R}_\varrho\left(\Phi_\varepsilon^f\right)(x) \leq 1$ for all $x \in \mathbb{R}^d$

Recall

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

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 $0 \leq \mathrm{R}_\varrho\left(\Phi_\varepsilon^H\right) \leq 1$ and $\left|H(x) - \mathrm{R}_\varrho\left(\Phi_\varepsilon^H\right)(x)\right| \leq \chi_{[0,\varepsilon]\times\mathbb{R}^{d-1}}(x)$ for all $x \in \mathbb{R}^d$ Moreover, $\left\|H - \mathrm{R}_\varrho\left(\Phi_\varepsilon^H\right)\right\|_{L^p([-1/2,1/2]^d)} \leq \varepsilon^{1/p}$ for all $p \in (0,\infty)$

[Proof of Lemma 3.4] Step 1: Approximation of γ

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

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

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

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

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

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

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

Step 2: Approximation of heaviside function

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

$$\mathrm{R}_\varrho\left(\Phi_\varepsilon^{\widetilde{\gamma}}\right)(x) = (x_1 + \gamma_\varepsilon(x_2, \ldots, x_d), x_2, \ldots, x_d) \quad \forall x \in \mathbb{R}^d$$

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

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

and $0 \leq \mathrm{R}_\varrho\left(\Phi_{\varepsilon'}^H\right) \leq 1$. 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 $s' := 2 + 3\lceil p \rceil$.

Step 3: Combining NNs from Step 1 and Step 2

Now we concatenate two networks above to get $\Phi^H_{\varepsilon'} \odot \Phi^{\widetilde{\gamma}}_{\varepsilon}$ with at most $\widetilde{L} \leq 2 + L(\Phi^{\widetilde{\gamma}}_{\varepsilon}) \leq 14 + (1 + \lceil \log_2 \beta \rceil) \cdot (11 + \frac{2\beta}{d})$ layers and no more than $\widetilde{c} \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero (s'', ε)-quantized weights for suitable $s'' = s''(d, \beta, B, p) \in \mathbb{N}$. Clearly, $0 \leq \mathrm{R}_\varrho(\Phi^H_{\varepsilon'} \odot \Phi^{\widetilde{\gamma}}_{\varepsilon}) \leq 1$.
Now, we check that $\mathrm{R}_\varrho(\Phi^H_{\varepsilon'} \odot \Phi^{\widetilde{\gamma}}_{\varepsilon})$ approximates $f = H \circ \widetilde{\gamma}$ with L^p-error at most ε.
Now, observe that for $q := 1 + \frac{1}{p} \geq max\{1, \frac{1}{p}\}$,

$$\left\| H \circ \widetilde{\gamma} - \mathrm{R}_\varrho(\Phi^H_{\varepsilon'} \odot \Phi^{\widetilde{\gamma}}_{\varepsilon}) \right\|_{L^p} = \left\| H \circ \widetilde{\gamma} - \mathrm{R}_\varrho(\Phi^H_{\varepsilon'}) \circ \mathrm{R}_\varrho(\Phi^{\widetilde{\gamma}}_{\varepsilon}) \right\|_{L^p}$$

$$\leq 2^q \cdot \max\left\{ \left\| H \circ \widetilde{\gamma} - H \circ \mathrm{R}_\varrho(\Phi^{\widetilde{\gamma}}_{\varepsilon}) \right\|_{L^p}, \left\| H \circ \mathrm{R}_\varrho(\Phi^{\widetilde{\gamma}}_{\varepsilon}) - \mathrm{R}_\varrho(\Phi^H_{\varepsilon'}) \circ \mathrm{R}_\varrho(\Phi^{\widetilde{\gamma}}_{\varepsilon})) \right\|_{L^p} \right\}$$

$$=: 2^q \cdot \max\{I, II\}$$

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

Step 4: Bounding term I

We first bound I from above.

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

Then, since

$$
\begin{aligned}
\chi_{\widetilde{\gamma}_1 \geq 0, \mathrm{R}_\varrho(\Phi^H_{\varepsilon'})_1 < 0} = 1 &\iff x_1 + \gamma(x_2, \dots, x_d) \geq 0 \text{ and } x_1 + \gamma_\varepsilon(x_2, \dots, x_d) < 0 \\
&\iff x_1 \in [-\gamma(x_2, \dots, x_d), -\gamma_\varepsilon(x_2, \dots, x_d))
\end{aligned}
$$

we have

$$
\int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{\widetilde{\gamma}_1 \geq 0, \mathrm{R}_\varrho(\Phi^H_{\varepsilon'})_1 < 0}(x_1, \dots, x_d) dx_1 \leq \max\{0, \gamma(x_2, \dots, x_d) - \gamma_\varepsilon(x_2, \dots, x_d)\}
$$

Step 5: Bounding term *II*

By similar reasoning, we also have

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{\tilde{\gamma}_1<0,\mathrm{R}_\varrho(\Phi^H_{\varepsilon'})_1\geq 0}(x_1,\ldots,x_d)dx_1 \leq \max\{0,\gamma_\varepsilon(x_2,\ldots,x_d)-\gamma(x_2,\ldots,x_d)\}$$

Hence,

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

This shows $2^q\cdot I\leq\frac{\varepsilon}{2}$.
Now we bound *II*. By construction of $\mathrm{R}_\varrho(\Phi^H_{\varepsilon'})$, we have $|H(x)-\mathrm{R}_\varrho(\Phi^H_{\varepsilon'})(x)|\leq \chi_{[0,\varepsilon']\times\mathbb{R}^{d-1}}(x)\leq\chi_{[0,\frac{1}{2}(\frac{\varepsilon}{4})^p]\times\mathbb{R}^{d-1}}(x)\ \forall x\in\mathbb{R}^d$.

Step 6: Combining Step 4 and Step 5

Therefore, we can write

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

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

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

Finally, noting that we can change coordinates by applying a permutation matrix to A_1 for any neural network $\Phi = ((A_1, b_1), \ldots, (A_L, b_L))$ 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. □

Theorem 3

Theorem 3: 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

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

Recall

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

[Proof] Step 1: Gridifying + Local Approximation

For $\lambda = (\lambda_1, \ldots, \lambda_d) \in \{1, \ldots, 2^r\}^d$, 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 $[-1/2, 1/2]^d = \bigcup_{\lambda \in \{1,\ldots,2^r\}^d} I_\lambda$, and $I_\lambda \subset \bar{B}_{2^{-r}}^{\|\cdot\|_{l^\infty}}(x)$ for all $x \in I_\lambda$.
Then, by definition of $\mathcal{K}_{r,\beta,d,B}$, for each $\lambda \in \{1, \ldots, 2^r\}^d$, $\exists f_\lambda \in \mathcal{HF}_{\beta,d,B}$ 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

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

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 $L_1 = L_1(d, \beta)$, $c_1 = c_1(d, \beta, B, r, p) > 0$, $s_1 = s_1(d, \beta, B, r, p) \in \mathbb{N}$

Step 2: Parallelizing and combining via "cutoff"

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

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

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

$$\|\mathrm{R}_\varrho(\Psi) - \chi_K\|_{L^p} = \left\| \mathrm{R}_\varrho(\Psi) - \sum_{I \in \{1,\ldots,2^{rd}\}} \chi_{I_{\lambda_I}} f_{\lambda_I} \right\|_{L^p}$$

$$\leq 2^q \left\| \mathrm{R}_\varrho(\Psi) - \sum_{I \in \{1,\ldots,2^r\}} \chi_{I_{\lambda_I}} [\mathrm{R}_\varrho(\Phi)]_I \right\|_{L^p} + 2^q \left\| \sum_{I \in \{1,\ldots,2^{rd}\}} \chi_{I_{\lambda_I}} \left([\mathrm{R}_\varrho(\Phi)]_I - f_{\lambda_I} \right) \right\|_{L^p}$$

$$\leq \frac{\varepsilon}{2} + 2^{q+rdq} \cdot \max_{I \in \{1,\ldots,2^{rd}\}} \left\{ \left\| \mathrm{R}_\varrho(\Phi_\varepsilon^{\lambda_I}) - f_{\lambda_I} \right\|_{L^p} \right\} \leq \varepsilon$$

Step 3: Bounding network output and check complexity

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 $(\max\{s_0, s_1\}, \varepsilon/(2^{1+q}))$-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 modulo multiplicative constants. □

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 $f \in \mathcal{E}^p_{r,\beta,d,B}$ there is a neural network Φ^f_ε 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

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

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

[Proof] Step 1: Apply THM 3.5 to indicator function

Let $q := \max\left\{1, p^{-1}\right\}, \varepsilon \in (0, 1/2)$ 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

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

Step 2: Apply THM 3.1 to smooth function

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

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

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.

Step 3: Construct multiplication network for Φ_ε^K and Φ_ε^g

Third, apply Lemma A.3 with $\theta = p(d-1)/\beta = d/\beta' \geq d/\beta_0$, $L_3^{(0)} := 1 + \lfloor \beta_0/(2d) \rfloor$ instead of L, $\frac{1}{3}\frac{1}{2^q}\varepsilon$ instead of ε and $M = \lceil B \rceil$ to obtain a network $\widetilde{\times}$ with at most

$$L_3 \leq 2 \cdot L_3^{(0)} + 8 \leq 10 + \beta_0/d$$

layers and at most $c_3 \cdot \varepsilon^{-\theta} = c_3 \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero, (s_3, ε)-quantized weights such that

$$|xy - \mathrm{R}_\varrho(\widetilde{\times})(x,y)| \leq \frac{\varepsilon}{3 \cdot 2^q} \quad \text{for all } x, y \in [-\lceil B \rceil, \lceil B \rceil]$$

Now set $\Psi_\varepsilon^f := \widetilde{x} \odot P\left(\Phi_\varepsilon^K, \Phi_\varepsilon^g\right)$. By properties of concatenation, Ψ_ε^f has at most

$$\max\{L_1, L_2\} + L_3 \leq 32 + (1 + \lceil \log_2 \beta_0 \rceil) \cdot (11 + 3\beta_0/d)$$

layers and $c_4 \cdot \varepsilon^{-p(d-1)/\beta}$ nonzero, $(\max\{s_1, s_2, s_3\}, \varepsilon)$-quantized weights for $c_4 = c_4(d, \beta, r, p, B) > 0$.

Step 4: Check the required approximation accuracy

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

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

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