Approximation Spaces and Deep Learning

Jake Lai* Supervisor Philipp Harms†

Abstract

Deep learning architectures underlie the most sophisticated and effective artificial intelligence applications today. While artificial neural networks were implemented as early as 1958 by Rosenblatt and known to possess the universal approximation property in the arbitrary width case since 1989 (due to Cybenko & others), mathematical explanations of their overperformance are still works in progress.

We define approximation schemes and approximation spaces, and present some fundamental results. Based on recent work by Gribonval et al., we relate approximation spaces of neural networks to the classical Besov spaces via embeddings in both directions, demonstrating the expressivity of deep neural networks.

Keywords: approximation spaces, neural networks, deep learning, Besov spaces, functional analysis, approximation theory

1 Introduction

In the span of a few years starting in 2011, the field of artificial intelligence experienced a remarkable revival, fuelled by unprecedented access to large amounts of data (colloquially, "Big Data") and deep learning. While artificial neural networks — the structures that deep learning techniques are based on — have been studied since at least 1943 by McCulloch and Pitts [6] and 1958 by Rosenblatt [9], they waxed and waned in popularity. Several breakthroughs, most prominently the success of a deep convolutional network in ImageNet's 2012 Challenge [5], have ushered in a new wave of artificial intelligence research focusing on deep learning. These have resulted in recent achievements, such as the text-to-image generative models DALL·E and Imagen, and the transformer-based language models GPT.

Deep learning, being a subset of the wider class of machine learning techniques, seems amenable to mathematical analysis. Indeed, some results have been es-

*Nanyang Business School, Nanyang Technological University. lair0004@e.ntu.edu.sg

†Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University. philipp.harms@ntu.edu.sg

tablished, such as the universal approximation property by Cybenko [1], Hornik et al. [4], and others. However, traditional statistical learning theory predicts that the high complexity of deep learning models should lead to poor generalisation performance; while optimisation theory predicts that solving very high-dimensional nonconvex optimisation problems, i.e., deep learning, should be intractable in general. This stands in contradiction to the observed performance of deep learning in practice. In general, many aspects of deep learning have yet to be satisfactorily captured by mathematical explanations; but a mathematical theory of deep learning is emerging.

Following recent work by [3], we recall the framework of approximation schemes and approximation spaces from functional analysis, including relevant fundamental results. We then use this formalism to provide embeddings in both directions of: 1. approximation spaces of functions representable by neural networks of a given number of weights and layers; and 2. the classical Besov spaces. This establishes the expressive capabilities of neural networks based on their complexity, i.e., number of weights and layers.

2 Theory of Approximation Spaces

In the following section, we recall the concept of approximation spaces as elaborated by [8].

2.1 Definitions

Definition 2.1. An *approximation scheme* (X, Σ) is a quasi-Banach space X with a sequence Σ of subsets $\Sigma_1 \subset \Sigma_2 \subset \cdots \subset X$ such that

1. $\lambda\Sigma_n \subset \Sigma_n$; and
2. $\Sigma_m + \Sigma_n \subset \Sigma_{m+n}$

for all nonnegative integers m, n. Fix $\Sigma_0 = \{0\}$.

Let the *error of best approximation* of f from a subset Γ be defined by $E_X(f, \Gamma) = \inf_{g \in \Gamma} \|f - g\|_X$.

Proposition 2.2. *Let (X, Σ) be an approximation scheme. Then for $f, g \in X$:*

1. $\|f\|_X = E_X(f, \Sigma_0) \geq E_X(f, \Sigma_1) \geq \cdots \geq 0$;
2. $E_X(\lambda f, \Sigma_n) = |\lambda| E_X(f, \Sigma_n)$; *and*
3. $E_X(f + g, \Sigma_{m+n}) \leq c_X(E_X(f, \Sigma_m) + E_X(g, \Sigma_n))$.

Definition 2.3. For an approximation scheme (X, Σ), the *approximation space* X_q^α consists of all $f \in X$ such that the sequence $\left(n^{\alpha - 1/q} E_X(f, \Sigma_{n-1})\right)$ is in ℓ^q, where n is a positive integer. Define

$$\|f\|_{X_q^\alpha} = \left\| \left(n^{\alpha - 1/q} E_X(f, \Sigma_{n-1}) \right) \right\|_{\ell^q}$$

for $f \in X_q^\alpha$.

Proposition 2.4. *X_q^α is a quasi-Banach space. Additionally, if $\alpha > \beta$, then $X_q^\alpha \hookrightarrow X_q^\beta$; and if $q \le s$, then $X_q^\alpha \hookrightarrow X_s^\beta$ for any α, β.*

Proposition 2.5. *Let $f \in X$. f is in X_q^α if and only if $\left(2^{k\alpha} E_X(f, \Sigma_{2^k-1})\right) \in \ell^q$. In addition, $\|f\|_{X_q^\alpha}^{\exp} = \|\left(2^{k\alpha} E_X(f, \Sigma_{2^k-1})\right)\|_{\ell^q}$ defines an equivalent quasi-norm on X_q^α.*

Proof. We prove the first statement. Note that $\frac{1}{2^k} + \frac{1}{2^k+1} + \cdots + \frac{1}{2^{k+1}-1} \ge \ln 2$. Hence, if $f \in X_q^\alpha$,

$$\begin{aligned}
\sum_{k=0}^{\infty} \left[2^{(k+1)\alpha} E_X(f, \Sigma_{2^{k+1}-1})\right]^q &\le \frac{2^{\alpha q}}{\ln 2} \sum_{k=0}^{\infty} 2^{k\alpha q} E_X(f, \Sigma_{2^{k+1}-1})^q \sum_{n=2^k}^{2^{k+1}-1} \frac{1}{n} \\
&\le c \sum_{k=0}^{\infty} E_X(f, \Sigma_{2^{k+1}-1})^q \sum_{n=2^k}^{2^{k+1}-1} n^{\alpha q - 1} \\
&\le c \sum_{k=0}^{\infty} \sum_{n=2^k}^{2^{k+1}-1} n^{\alpha q - 1} E_X(f, \Sigma_{n-1})^q \\
&\le c \sum_{n=1}^{\infty} \left[n^{\alpha q - 1} E_X(f, \Sigma_{n-1})\right]^q .
\end{aligned}$$

The converse is straightforward. □

2.2 Theorems

We now turn to some fundamental theorems on approximation spaces.

Theorem 2.6 (Representation Theorem)**.** *Let (X, Σ) be an approximation scheme. Then $f \in X$ is in X_q^α if and only if there exist $f_k \in \Sigma_{2^k}$ such that $f = \sum_{k=0}^{\infty} f_k$ (the so-called "representation") and $\left(2^{k\alpha} \|f_k\|_X\right) \in \ell^q$. Moreover, define*

$$\|f\|_{X_q^\alpha}^{\mathrm{rep}} = \inf \left\| \left(2^{k\alpha} \|f_k\|_X\right) \right\|_{\ell^q},$$

where the infimum is taken over all possible representations. $\|\cdot\|_{X_q^\alpha}^{\mathrm{rep}}$ is an equivalent quasi-norm on X_q^α.

Proof. Let $f \in X_q^\alpha$. For each k, choose $f_k^* \in \Sigma_{2^k}$ such that $\|f - f_k^*\|_X \le 2E_X(f, \Sigma_{2^k})$, and set $f_0 = f_1 = 0$ and $f_k = f_{k-1}^* - f_{k-2}^*$. We have $f_k \in \Sigma_{2^{k-1}+2^{k-2}} \subset \Sigma_{2^k}$, and

$$f = \lim_{k\to\infty} f_k^* = \sum_{k=0}^{\infty} f_k.$$

Observe that

$$\begin{aligned}\|f_k\|_X &\le c_X(\|f - f^*_{k-1}\|_X + \|f - f^*_{k-2}\|_X)\\ &\le 4c_X E_X(f, \Sigma_{2^{k-2}-1}).\end{aligned}$$

Applying 2.5, we find that $\left(2^{k\alpha}\|f_k\|_X\right) \in \ell^q$ and

$$\|f\|^{\rm rep}_{X^\alpha_q} \le 2^{2\alpha+2} c_X \|f\|^{\rm exp}_{X^\alpha_q}.$$

Call a quasi-norm $\|\cdot\|$ a *p-norm* if $\|f+g\| \le \|f\|^p + \|g\|^p$ for all f, g. Without loss of generality, assume that $\|\cdot\|_X$ is a p-norm, where $0 < p < q$. If f has a representation $f = \sum_{k=0}^\infty f_k$ satisfying $f_k \in \Sigma_{2^k}$ and $\left(2^{k\alpha}\|f_k\|_X\right) \in \ell^q$, then since $\sum_{k=0}^{n-1} f_k \in \Sigma_{2^n-1}$, we have that

$$E_X(f, \Sigma_{2^n-1})^p \le \left\| f - \sum_{k=0}^{n-1} f_k \right\|^p_X \le \sum_{k=n}^\infty \|f_k\|^p_X.$$

When $0 < q < \infty$, fix $r = q/p$ and β such that $0 < \beta < \alpha p$. Apply Hölder's inequality to obtain

$$\begin{aligned}\sum_{n=0}^\infty (2^{n\alpha} E_X(f, \Sigma_{2^n-1}))^q &\le \sum_{n=0}^\infty 2^{n\alpha q} \left(\sum_{k=n}^\infty 2^{-k\beta} 2^{k\beta} \|f_k\|^p_X \right)^r\\ &\le \sum_{n=0}^\infty 2^{n\alpha q} \left(\sum_{k=n}^\infty 2^{-k\beta r'} \right)^{r/r'} \left(\sum_{k=n}^\infty 2^{k\beta r} \|f_k\|^q_X \right)\\ &\le c_1 \sum_{n=0}^\infty 2^{n(\alpha q - \beta r)} \sum_{k=n}^\infty 2^{k\beta r} \|f_k\|^q_X\\ &\le c_1 \sum_{k=1}^\infty 2^{k\beta r} \|f_k\|^q_X \sum_{n=0}^k 2^{n(\alpha q - \beta r)}\\ &\le c_2 \sum_{k=1}^\infty (2^{k\alpha}\|f_k\|_X)^q < \infty.\end{aligned}$$

Thus, $\|f\|^{\rm exp}_{X^\alpha_q} \le c \left\| \left(2^{k\alpha}\|f_k\|_X\right) \right\|_{\ell^q}$, and by 2.5, $f \in X^\alpha_q$ and $\|f\|^{\rm exp}_{X^\alpha_q} \le c\|f\|^{\rm rep}_{X^\alpha_q}$. A similar argument follows for $q = \infty$. □

Corollary 2.7. *If $0 < q < \infty$, then the linear subset $\Sigma_\infty = \bigcup_{n=1}^\infty \Sigma_n$ is dense in X^α_q.*

Call an approximation scheme (X, Σ) *linear* if there exists a uniformly bounded sequence P_n of linear projections mapping X onto Σ_n.

Theorem 2.8 (Linear Representation Theorem)**.** *Let (X, Σ) be a linear approximation scheme. Then $f \in X$ is in X_q^α if and only if $\left(2^{k\alpha}\|f_k\|_X\right) \in \ell^q$, where $f_k = (P_{2^{k+1}-1} - P_{2^k-1})f$. In particular, we have the linear representation*

$$f = \sum_{k=0}^{\infty} f_k,$$

and

$$\|f\|_{X_q^\alpha}^{\text{lin}} \leq \left\|\left(2^{k\alpha}\|f_k\|_X\right)\right\|_{\ell^q}$$

defines an equivalent quasi-norm on X_q^α.

Theorem 2.9 (Reiteration Theorem)**.** *Let (X, Σ) be an approximation scheme. Then $(X_q^\alpha)_s^\beta = X_s^{\alpha+\beta}$. (In other words, we can iteratively "construct" approximation spaces.)*

Proof. If $f \in (X_q^\alpha)_s^\beta$, then $\left(2^{k\beta}E_{X_q^\alpha}(f, \Sigma_{2^k-1})\right) \in \ell^s$. By Lemma 1 of [8, Theorem 3.2], which states that there exists a constant $c > 0$ such that $n^\alpha E_X(f, \Sigma_{2n-2}) \leq cE_{X_q^\alpha}(f, \Sigma_{n-1})$ for $n = 1, 2, \ldots$, we obtain $\left(2^{k(\alpha+\beta)}E_X(f, \Sigma_{2^k-1})\right) \in \ell^s$. Thus, $f \in X_s^{\alpha+\beta}$.

Conversely: if $f \in X_s^{\alpha+\beta}$, then we have a representation $f = \sum_{k=0}^\infty f_k$ with $f_k \in \Sigma_{2^k}$ and $\left(2^{k(\alpha+\beta)}\|f_k\|_X\right) \in \ell^s$. By Lemma 2 of [8, Theorem 3.2], which states that there exists a constant $c > 0$ such that $\|f\|_{X_q^\alpha} \leq cn^\alpha\|f\|_X$ for all $f \in \Sigma_n$ and $n = 1, 2, \ldots$, we have that $\left(2^{k\beta}\|f_k\|_{X_q^\alpha}\right) \in \ell^s$. Therefore, $f \in (X_q^\alpha)_s^\beta$. □

Let $\mathcal{L}(X, Y)$ be the (quasi-Banach) space of bounded linear operators from quasi-Banach spaces X to Y with the operator quasi-norm.

Theorem 2.10 (Transformation Theorem)**.** *Let (X, A) and (Y, B) be approximation schemes, and $T \in \mathcal{L}(X, Y)$. Suppose there exist positive constants c, β such that $TA_m \subset B_n$ when $n \geq cm^\beta$. Then $T(X_q^{\alpha\beta} \subset Y_q^\alpha$ for all α, q.*

Proof. Consider $\beta \geq 1$. Let $N_m = \{n \in \mathbb{N} : c(m-1)^\beta + 1 \leq n < cm^\beta + 1\}$ for $m = 1, 2, \ldots$ — these partition the positive integers. It may be seen that:

1. $|N_m| \leq c_1 m^{\beta-1}$;
2. $n^{\alpha-1/q} \leq c_2 m^{\beta(\alpha-1/q)}$ for $n \in N_m$; and
3. $E_Y(Tf, B_{n-1} \leq \|T\|_{\mathcal{L}} E_X(f, A_{m-1})$ for $f \in X$ and $n \in N_m$.

Hence, if $0 < q < \infty$, we have

$$\begin{aligned}\|Tf\|_{Y_q^\alpha} &= \left(\sum_{m=1}^{\infty}\sum_{n\in N_m}[n^{\alpha-1/q}E_Y(Tf,B_{n-1})]^q\right)^{1/q}\\ &\le \left(\sum_{m=1}^{\infty} c_1 m^{\beta-1}[c_2 m^{\beta(\alpha-1/q)}\|T\|_{\mathcal{L}}E_X(f,A_{m-1})]^q\right)^{1/q}\\ &\le c\|f\|_{X_q^\alpha}\end{aligned}$$

The case $0 < \beta < 1$ follows similarly, as do the cases where $q = \infty$. □

We omit the proofs of the following theorems, which can be found in [8].

Theorem 2.11 (Embedding Theorem)**.** *Let X and Y be quasi-Banach spaces continuously embedded into some linear topological (Hausdorff) space, and let (X,Σ) and (Y,Σ) be approimation schemes with the same sequence of subsets Σ. Suppose there exist positive constants c,β such that $\|f\|_Y \le cn^\beta\|f\|_X$ for all $f\in\Sigma_n$ and $n = 1,2,\ldots$. Then $X_q^{\alpha+\beta}\subset Y_q^\alpha$. In particular, if Y can be equipped with a p-norm with $0 < p \le 1$, then $X_p^\beta \subset Y$.*

Theorem 2.12 (Composition Theorem)**.** *Let (X,A), (Y,B), and (Z,C) be approximation schemes. Let $M : X\times Y\to Z$ be a bounded bilinear map. Suppose $M(A_n,Y)$ and $M(X,B_n)$ are both contained in C_n. Then $M(X_p^\alpha,Y_q^\beta)\subset Z_r^{\alpha+\beta}$, where $1/r = 1/p + 1/q$.*

3 Approximation Spaces of Deep Neural Networks

We can now treat deep neural networks with the formalism of approximation spaces, following [3]. We only consider "strict" neural networks below, omitting discussion of generalised neural networks encompassing networks with skip connections like Residual Networks. (It turns out that the approximation spaces of strict and generalised networks coincide for activation functions that can represent the identity [3, Theorem 3.8].)

3.1 Definitions

Definition 3.1. A *neural network* Φ with *activation function* or *nonlinearity* $\varrho : \mathbb{R}\to\mathbb{R}$ is an ordered tuple $((T_1,\alpha_1),\ldots,(T_L,\alpha_L))$, where for each $l = 1,2,\ldots,L$, N_l is a positive integer, $T_l : \mathbb{R}^{N_{l-1}}\to\mathbb{R}^{N_l}$ is an affine map, $\alpha_l : \mathbb{R}^{N_l}\to\mathbb{R}^{N_l}$ applies ϱ coordinate-wise for $1\le l < L$, i.e., $(x_1,\ldots,x_{N_l}) \overset{\alpha_l}{\longmapsto} (\varrho(x_1),\ldots,\varrho(x_{N_l}))$, and $\alpha_L = \mathrm{id}_{\mathbb{R}^{N_L}}$. Such a neural network is said to have a *depth* of $L(\Phi) = L$ "layers", each of width N_l. The *number of weights* is given by $W(\Phi) = \sum_{l=1}^{L}\|T_l\|_{\ell^0}$, where $\|T\|_{\ell^0}$ counts the nonzero entries in the matrix A if

the affine map $Tx = Ax + b$ (we say A is the "matrix part" of T). The *number of hidden neurons* is given by $N(\Phi) = \sum_{l=1}^{L-1} N_l$.

Definition 3.2. The *realisation* $\mathcal{R}(\Phi)$ of a neural network $((T_1, \alpha_1), \ldots, (T_L, \alpha_L))$ is the function

$$\mathcal{R}(\Phi) = \alpha_L \circ T_L \circ \cdots \circ \alpha_1 \circ T_1.$$

Definition 3.3. Let L be a positive integer (possibly ∞), W and N be nonnegative integers (also possibly ∞), and $\Omega \subset \mathbb{R}^d$ nonempty. $\mathcal{NN}_{W,L,N}^{\varrho,d,k}$ is the set of all networks Φ with activation function ϱ, *input dimension* $N_0 = d$, *output dimension* $N_L = k$, number of weights $W(\Phi) \leq W$, depth $L(\Phi) \leq L$, and number of hidden neurons $N(\Phi) \leq N$. $\mathtt{NN}_{W,L,N}^{\varrho,d,k}$ is the set of all functions representable by networks in $\mathcal{NN}_{W,L,N}^{\varrho,d,k}$; i.e., $\mathtt{NN}_{W,L,N}^{\varrho,d,k} = \{\mathcal{R}(\Phi) : \Phi \in \mathcal{NN}_{W,L,N}^{\varrho,d,k}$. Finally, $\mathtt{NN}_{W,L,N}^{\varrho,d,k}(\Omega)$ is the set of all such functions restricted to Ω.

Note. Our set of strict neural networks $\mathcal{NN}_{W,L,N}^{\varrho,d,k}$ corresponds to the set $\mathcal{SNN}_{W,L,N}^{\varrho,d,k}$ in [3], which uses $\mathcal{NN}_{W,L,N}^{\varrho,d,k}$ to refer to generalised networks instead. (Similarly for $\mathtt{NN}_{W,L,N}^{\varrho,d,k}$ and $\mathtt{SNN}_{W,L,N}^{\varrho,d,k}$.)

Definition 3.4. A *depth-growth function* is a nondecreasing function $\mathcal{F} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$. $\mathcal{F}$ is *dominated* by $\mathcal{F}'$ (written $\mathcal{F} \preceq \mathcal{F}'$) if, for large enough n, there exists c a positive integer such that $\mathcal{F}(n) \leq \mathcal{F}'(cn)$. $\mathcal{F}$ is equivalent to $\mathcal{F}'$ (written $\mathcal{F} \sim \mathcal{F}'$) if $\mathcal{F} \preceq \mathcal{F}' \preceq \mathcal{F}$.

Definition 3.5. Let ϱ be an activation function, $\mathcal{F}$ a depth-growth function, Ω a subset of $\mathbb{R}^d$, and X a quasi-Banach space consisting of functions $f : \Omega \to \mathbb{R}^k$. Define $\mathtt{W}_n(X, \varrho, \mathcal{F}) = \mathtt{NN}_{n,\mathcal{F}(n),\infty}^{\varrho,d,k}(\Omega) \cap X$, setting $\mathtt{W}_0(X, \varrho, \mathcal{F}) = \{0\}$. (The depth-growth function controls how the depth of our networks grows with the number of weights n.)

Proposition 3.6. *Let $\Sigma_n = \mathtt{W}_n(X, \varrho, \mathcal{F})$. Then (X, Σ) is in fact an approximation scheme. Denote the associated approximation spaces by $W_q^\alpha(X, \varrho, \mathcal{F}) = X_q^\alpha$.*

Proof. $\Sigma_0 = \mathtt{W}_0(X, \varrho, \mathcal{F}) = \{0\}$ by definition. Since $\mathtt{NN}_{W,L,\infty}^{\varrho,d,k}(\Omega) \subset \mathtt{NN}_{W+1,L',\infty}^{\varrho,d,k}(\Omega)$ whenever $L \leq L'$, and depth-growth functions are by definition nondecreasing, we immediately have

$$\mathtt{W}_n(X, \varrho, \mathcal{F}) = \mathtt{NN}_{n,\mathcal{F}(n),\infty}^{\varrho,d,k}(\Omega) \subset \mathtt{NN}_{n+1,\mathcal{F}(n+1),\infty}^{\varrho,d,k}(\Omega) = \mathtt{W}_{n+1}(X, \varrho, \mathcal{F}).$$

Let $f \in \mathtt{NN}_{W,L,\infty}^{\varrho,d,k}(\Omega)$, and suppose $f = \mathcal{R}(\Phi) = \alpha_L \circ T_L \circ \cdots \circ \alpha_1 \circ T_1$. Simply take $\lambda f = \mathcal{R}(\Phi') = \alpha_L \circ (\lambda T_L) \circ \cdots \circ \alpha_1 \circ T_1$ (remembering that $\alpha_L = \mathrm{id}_{\mathbb{R}^k}$). This shows that $\lambda \mathtt{W}_n(X, \varrho, \mathcal{F}) \subset \mathtt{W}_n(X, \varrho, \mathcal{F})$.

Finally, without loss of generality, suppose $m \leq n$. Let $f \in \mathtt{NN}_{m,L,\infty}^{\varrho,d,k}(\Omega)$ and $g \in \mathtt{NN}_{n,L,\infty}^{\varrho,d,k}(\Omega)$, and suppose that $f = \mathcal{R}(\Phi_1) = \alpha_L \circ T_L \circ \cdots \circ \alpha_1 \circ T_1$ and $g = \mathcal{R}(\Phi_2) = \beta_L \circ S_L \circ \cdots \circ \beta_1 \circ S_1$. One may check that $f + g = \mathcal{R}(\Phi) =$

$\theta_L \circ R_L \circ \cdots \circ \theta_1 \circ R_1$, where $R_l : \mathbb{R}^{M_{l-1}+N_{l-1}} \to \mathbb{R}^{M_l+N_l}; (x,y) \mapsto (T_l x, S_l y)$ and $R_L : \mathbb{R}^{M_{L-1}+N_{L-1}} \to \mathbb{R}^k; (x,y) \mapsto T_L x + S_L y)$; while $\theta_l : \mathbb{R}^{M_l+N_l} \to \mathbb{R}^{M_l+N_l}; (x,y) \mapsto (\alpha_l x, \beta_l y)$ and $\theta_L = \mathrm{id}_{\mathbb{R}^k}$; for $1 \leq l < L$. The number of weights of the new network Φ is $W(\Phi) = W(\Phi_1) + W(\Phi_2)$ — if $R_l x = A_l x + b_l$, A_l has the matrix part of T_l in the top-left block, the matrix part of S_l in the bottom-right block, and zeroes everywhere else; A_L has the matrix part of T_L on the left and the matrix part of S_L on the right. Thus, taking $L = \max\{\mathcal{F}(m), \mathcal{F}(n)\} \leq \mathcal{F}(m+n)$ shows that $\mathtt{W}_m(X, \varrho, \mathcal{F}) + \mathtt{W}_n(X, \varrho, \mathcal{F}) \subset \mathtt{W}_{m+n}(X, \varrho, \mathcal{F})$. □

Note. [3] places an additional constraint on approximation schemes; namely that Σ_∞ is dense in X. It is seen that the networks we wish to address satisfy the necessary measure-theoretic requirements to meet this constraint.

3.2 ReLU-networks and related networks

Write $\varrho_1 : \mathbb{R} \to \mathbb{R}^+$ for the activation function $x \overset{\varrho_1}{\longmapsto} \max\{0, x\}$; ϱ_1 is known as the *rectified linear unit* or *ReLU*. Write ϱ_r for the r^{th} power of ϱ_1; i.e., $x \overset{\varrho_r}{\longmapsto} \varrho_1(x)^r$. We will be concerning ourselves primarily with ϱ_r-networks, as the ReLU is one of the most commonly used nonlinearities for current deep learning practice.

Definition 3.7. Let $I \subset \mathbb{R}$ be an interval. The set of *piecewise polynomials* of degree at most r with at most n pieces (or $n-1$ breakpoints) is denoted $\mathtt{PPoly}_n^r(I)$; write $\mathtt{PPoly}^r(I) = \bigcup_{n\in\mathbb{N}} \mathtt{PPoly}_n^r(I)$. The set of *free-knot splines* of degree at most r with at most n pieces (or $n-1$ breakpoints) is defined as $\mathtt{Spline}_n^r(I) = \mathtt{PPoly}_n^r(I) \cap C^{r-1}(I)$; again, write $\mathtt{Spline}^r(I) = \bigcup_{n\in\mathbb{N}} \mathtt{Spline}_n^r(I)$.

Call a domain $\Omega \subset \mathbb{R}^d$ *admissible* if it is Borel-measurable with nonzero measure.

Theorem 3.8. *Consider the space $X = L^p(\Omega; \mathbb{R}^k)$, where $0 < p \leq \infty$ and $\Omega \subset \mathbb{R}^d$ is an admissible domain. Let $\mathcal{F}$ be a depth-growth function.*

1. *If $\varrho \in \mathtt{PPoly}^r(\mathbb{R})$, then*

$$W_q^\alpha(X, \varrho, \mathcal{F}) \hookrightarrow \begin{cases} W_q^\alpha(X, \varrho_r, \max\{\mathcal{F}+1, 2\}) & \text{if } d = 1 \\ W_q^\alpha(X, \varrho_r, \max\{\mathcal{F}+1, 3\}) & \text{if } d \geq 2 \end{cases}.$$

   Furthermore, if Ω is bounded or $r = 1$ or $\mathcal{F}+1 \preceq \mathcal{F}$, then $W_q^\alpha(X, \varrho, \mathcal{F}) \hookrightarrow W_q^\alpha(X, \varrho_r, \mathcal{F})$.

2. *If $\varrho \in \mathtt{Spline}^r(\mathbb{R})$ and Ω is bounded, then $\mathcal{F}+1 \preceq \mathcal{F}$, then $W_q^\alpha(X, \varrho, \mathcal{F}) = W_q^\alpha(X, \varrho_r, \mathcal{F})$ with equivalent norms.*

3. *For any positive integer s, we have $W_q^\alpha(X, \varrho_{r^s}, \mathcal{F}) \hookrightarrow W_q^\alpha(X, \varrho_r, s(\mathcal{F}-1))$.*

Corollary 3.9. *If $2\mathcal{F} \preceq \mathcal{F}$, then for all $r \geq 2$ we have*

$$W_q^\alpha(X, \varrho_1, \mathcal{F}) \hookrightarrow W_q^\alpha(X, \varrho_2, \mathcal{F}) = W_q^\alpha(X, \varrho_r, \mathcal{F}).$$

Theorem 3.10. *Let $X = L^p(\Omega; \mathbb{R}^k)$, where $0 < p, q \leq \infty$ and $\Omega \subset \mathbb{R}^d$ be measurable with interior nonempty. Let $\varrho \in \mathtt{PPoly}_n^r(\mathbb{R})$ be continuous and $\mathcal{F}$ be a depth-growth function with $\sup \mathcal{F} \geq 2$. Then $W_q^\alpha(X, \varrho, \mathcal{F}) \subsetneq X$. ($W_q^\alpha(X, \varrho, \mathcal{F})$ is a proper, i.e., nontrivial subset of X.)*

3.3 Limitations of bounded depth ReLU-networks

Denote the class of k-times continuously differentiable functions on Ω with compact support by $C_c^k(\Omega)$.

Theorem 3.11. *Let $\Omega \subset \mathbb{R}^d$ be open and admissible, $0 < p, q \leq \infty$, $X = L^p(\Omega)$, and L be a positive integer. If $C_c^3(\Omega) \cup W_q^\alpha(X, \varrho_1, L) \neq \{0\}$, then $\lfloor L/2 \rfloor \geq \alpha/2$.*

Taking the contrapositive, we see that ReLU-networks with bounded depth fail to contain any nonzero function $f \in C_c^3(\Omega)$.

Proof (sketch). Since $W_q^\alpha(X, \varrho_1, L) \subset W_\infty^\alpha(X, \varrho_1, L)$, we can consider $q = \infty$. Let $f \in C_c^3(\Omega)$ be nonzero. Choose $\Omega_0 = B_r(x_0) \subset \Omega$ so $f|_{\Omega_0}$ is not affine and, by [7, Proposition C.5], there is a constant $c_1 > 0$ such that $\|f - g\|_{L^p(\Omega_0)} \geq c_1 P^{-2}$ for each P-piecewise slice affine function g. There is a constant K so that $\mathtt{NN}_{W,L,\infty}^{\varrho_1,1,1} \subset \mathtt{PPoly}_{KW^{\lfloor L/2 \rfloor}}^1(\mathbb{R})$. Thus, each $g \in \mathtt{NN}_{W,L,\infty}^{\varrho_1,d,1}$ is P-piecewise slice affine with $P = KW^{\lfloor L/2 \rfloor}$.

Now let $f \in W_\infty^\alpha(X, \varrho_1, L)$. There is a constant $c_2 > 0$ such that for each positive integer n there is $g_n \in \mathtt{NN}_{n,L,\infty}^{\varrho_1,d,1}$ satisfying $\|f - g\|_{L^p(\Omega_0)} \leq \|f - g\|_X \leq c_2 n^{-\alpha}$.

Combining, we see that $K^{-2} c_1 n^{-2\lfloor L/2 \rfloor} \leq \|f - g\|_{L^p(\Omega_0)} \leq c_2 n^{-\alpha}$, so $\alpha \leq 2\lfloor L/2 \rfloor$. □

3.4 Relations to Besov spaces

We now study embeddings of Besov spaces into $W_q^\alpha(X, \varrho_r, L)$ — which we shall call *direct estimates*; and embeddings of $W_q^\alpha(X, \varrho_r, L)$ into Besov spaces — we will call these *inverse estimates*.

The *difference operator* is defined as $\Delta_h f(x) = f(x + h) - f(x)$, and

$$\Delta_h(f, x, \Omega) = \begin{cases} \Delta_h f(x) & \text{if } x, x + h, \ldots, x + rh \in \Omega \\ 0 & \text{otherwise} \end{cases}.$$

The *modulus of smoothness of order r* of $f \in L^p(\Omega)$ is

$$\omega_r(f, t)_p = \sup_{|h| \leq t} \|\Delta_h(f, \cdot, \Omega)^r\|_{L^p(\Omega)}.$$

Definition 3.12. [2, Section 2] Let $\Omega \subset \mathbb{R}^d$ be open, $\alpha > 0$, $0 < p \leq \infty$, and $1 \leq q < \infty$. The *Besov space* $B_{p,q}^s(\Omega)$ consists of all functions $f \in L^p(\Omega)$ such

that

$$|f|_{B^s_{p,q}(\Omega)} = \left(\int_0^\infty \left| t^{-s} \omega_r(f,t,\Omega)_p \right|^q \frac{dt}{t} \right)^{1/q} < \infty.$$

Define the norm on $B^s_{p,q}(\Omega)$ as such:

$$\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + |f|_{B^s_{p,q}(\Omega)}.$$

We omit the (spline-/wavelet-theoretic) proofs of the following estimates, which are laid out in [3].

Theorem 3.13. *Let Ω be a bounded Lipschitz domain with nonzero measure, and $X = L^p(\Omega)$. Let $\mathcal{F}$ be a depth-growth function.*

1. *If $d = 1$ and $L = \sup \mathcal{F}(n) \geq 2$, then for all $0 < p, q \leq \infty$ and $0 < s < r + \min\{1, 1/p\}$, we have*

$$B^s_{p,q} \hookrightarrow W^s_q(X, \varrho_r, \mathcal{F}).$$

2. *If $d > 1$ and $L = \sup \mathcal{F}(n) \geq 3$, then for all $0 < p, q \leq \infty$ and $0 < s < \frac{1}{d}(r_0 + \min\{1, 1/p\})$, we have*

$$B^{sd}_{p,q} \hookrightarrow W^s_q(X, \varrho_r, \mathcal{F}),$$

letting $r_0 = \begin{cases} r & \text{if } r \geq 2 \text{ and } L \geq 2 + 2\lceil \log_2 d \rceil \\ 0 & \text{otherwise} \end{cases}$.

The following theorem shows that $W^\alpha_q(X, \varrho_r, \mathcal{F})$ can fail to embed into Besov spaces if the approximation rate parameter α is too small.

Theorem 3.14. *Let $\Omega = (0,1)^d$ and $X = L^p(\Omega)$. Let $\mathcal{F}$ be a depth-growth function with $L = \sup \mathcal{F}(n) \geq 2$. For all $0 < \sigma, \tau, q \leq \infty$ and $\alpha, s > 0$, if $W^\alpha_q(X, \varrho_r, \mathcal{F}) \hookrightarrow B^s_{\sigma,\tau}(\Omega)$, then $\alpha \geq \lfloor L/2 \rfloor \cdot \min\{s, 2\}$.*

Theorem 3.15. *Let $\Omega = (0,1)$ and $X = L^p(\Omega)$ (the above hypotheses for $d = 1$). Let $\mathcal{F}$ be a depth-growth function with $L = \sup \mathcal{F}(n) < \infty$. Set $s = \alpha / \lfloor L/2 \rfloor$. When $q = 1/(s + 1/p)$, we have*

$$W^\alpha_q(X, \varrho_r, \mathcal{F}) \hookrightarrow B^s_{q,q}(\Omega).$$

In fact, $W^\alpha_q(X, \varrho_r, \mathcal{F})$ embeds into a real interpolation space of $L^p(\Omega)$ and the above Besov space for all $s > 0$ and $0 < \alpha < s\lfloor L/2 \rfloor$.

4 Conclusion

The language of approximation spaces can be a fruitful framework with which we may analyse the expressivity of a large class of neural networks.

It is proven in [3, Lemma 2.18] that the class of neural networks is closed under composition. In future work, we hope to further study approximation spaces of functions satisfying structural conditions like $\mathtt{W}_m(X, \varrho, \mathcal{F}) \circ \mathtt{W}_n(X, \varrho, \mathcal{F}) \subset \mathtt{W}_{m+n}(X, \varrho, \mathcal{F})$. By exploring the expanded approximation capacity that such compositional properties provide, we hope to compare neural networks to wavelets, splines, and other approximation techniques.

Acknowledgment

I would like to thank the faculty of the Division of Mathematical Sciences under the School of Physical and Mathematical Sciences, Nanyang Technological University for inspiring and encouraging me to push forward on my mathematical journey, with especial thanks to my supervisor Associate Professor Philipp Harms for his invaluable guidance and patience. Other faculty members I would like to thank for their time and support include Senior Lecturer Gary Royden Watson Greaves, Senior Lecturer Lim Kay Jin, and Associate Professor Xia Kelin.

I would also like to thank my fellow URECA participant Chan Joshua Juan Yin for our discussions, about mathematics or otherwise.

This research project was supported by Nanyang Technological University under the URECA Undergraduate Research Programme.

References

[1] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals and Systems* 2.4 (Dec. 1, 1989), pp. 303–314. DOI: 10.1007/BF02551274. URL: https://link.springer.com/article/10.1007/BF02551274.

[2] Ronald A. Devore and Robert C. Sharpley. "Besov Spaces on Domains in Rd". In: *Transactions of the American Mathematical Society* 335.2 (1993), pp. 843–864. DOI: 10.2307/2154408. URL: https://www.jstor.org/stable/2154408.

[3] Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. "Approximation Spaces of Deep Neural Networks". In: *Constructive Approximation* 55.1 (Feb. 1, 2022), pp. 259–367. DOI: 10.1007/s00365-021-09543-4. URL: https://link.springer.com/article/10.1007/s00365-021-09543-4.

[4] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (Jan. 1, 1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012. URL: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[6] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The Bulletin of Mathematical Biophysics* 5.4 (Dec. 1, 1943), pp. 115–133. DOI: 10.1007/BF02478259. URL: https://link.springer.com/article/10.1007/BF02478259.

[7] Philipp Petersen and Felix Voigtlaender. "Optimal approximation of piecewise smooth functions using deep ReLU neural networks". In: *Neural Networks* 108 (Dec. 2018), pp. 296–330. DOI: 10.1016/j.neunet.2018.08.019. URL: https://www.sciencedirect.com/science/article/abs/pii/S0893608018302454.

[8] Albrecht Pietsch. "Approximation spaces". In: *Journal of Approximation Theory* 32.2 (June 1, 1981), pp. 115–134. DOI: 10.1016/0021-9045(81)90109-X. URL: https://www.sciencedirect.com/science/article/pii/002190458190109X.

[9] Frank Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Fh0042519.