Chapter 13

Shape of neural network spaces

As we have seen in the previous chapter, the loss landscape of neural networks can be quite intricate and is typically not convex. In some sense, the reason for this is that we take the point of view of a map from the parameterization of a neural network. Let us consider a convex loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a sample $S = (\boldsymbol{x}_i, y_i)_{i=1}^m \in (\mathbb{R}^d \times \mathbb{R})^m$.

Then, for two neural networks Φ_1, Φ_2 and for $\alpha \in (0, 1)$ it holds that

$$\begin{aligned}\widehat{\mathcal{R}}_S(\alpha\Phi_1 + (1-\alpha)\Phi_2) &= \frac{1}{m}\sum_{i=1}^m \mathcal{L}(\alpha\Phi_1(\boldsymbol{x}_i) + (1-\alpha)\Phi_2(\boldsymbol{x}_i), y_i) \\ &\leq \frac{1}{m}\sum_{i=1}^m \alpha\mathcal{L}(\Phi_1(\boldsymbol{x}_i), y_i) + (1-\alpha)\mathcal{L}(\Phi_2(\boldsymbol{x}_i), y_i) \\ &= \alpha\widehat{\mathcal{R}}_S(\Phi_1) + (1-\alpha)\widehat{\mathcal{R}}_S(\Phi_2).\end{aligned}$$

Hence, the empirical risk is convex when considered as a map depending on the neural network functions rather then the neural network parameters. A convex function does not have spurious minima or saddle points. As a result, the issues from the previous section are avoided if we take the perspective of neural network sets.

So why do we not optimize over the sets of neural networks instead of the parameters? To understand this, we will now study the set of neural networks associated to a fixed architecture as a subset of other function spaces.

We start by investigating the realization map R_σ introduced in Definition 12.1. Concretely, we show in Section 13.1, that if σ is Lipschitz, then the set of neural networks is the image of $\mathcal{PN}(\mathcal{A}, \infty)$ under a locally Lipschitz map. We will use this fact to show in Section 13.2 that sets of neural networks are typically non-convex, and even have arbitrarily large holes. Finally, in Section 13.3, we study the extent to which there exist best approximations to arbitrary functions, in the set of neural networks. We will demonstrate that the lack of best approximations causes the weights of neural networks to grow infinitely during training.

13.1 Lipschitz parameterizations

In this section, we study the realization map R_σ. The main result is the following simplified version of [171, Proposition 4].

Proposition 13.1. *Let $\mathcal{A} = (d_0, d_1, \dots, d_{L+1}) \in \mathbb{N}^{L+2}$, let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be C_σ-Lipschitz continuous with $C_\sigma \geq 1$, let $|\sigma(x)| \leq C_\sigma |x|$ for all $x \in \mathbb{R}$, and let $B \geq 1$.*

Then, for all $\theta, \theta' \in \mathcal{PN}(\mathcal{A}, B)$,

$$\|R_\sigma(\theta) - R_\sigma(\theta')\|_{L^\infty([-1,1]^{d_0})} \leq (2C_\sigma B d_{\max})^L n_{\mathcal{A}} \|\theta - \theta'\|_\infty,$$

where $d_{\max} = \max_{\ell=0,\dots,L+1} d_\ell$ and $n_{\mathcal{A}} = \sum_{\ell=0}^{L} d_{\ell+1}(d_\ell + 1)$.

Proof. Let $\theta, \theta' \in \mathcal{PN}(\mathcal{A}, B)$ and define $\delta \coloneqq \|\theta - \theta'\|_\infty$. Repeatedly using the triangle inequality we find a sequence $(\theta_j)_{j=0}^{n_{\mathcal{A}}}$ such that $\theta_0 = \theta$, $\theta_{n_{\mathcal{A}}} = \theta'$, $\|\theta_j - \theta_{j+1}\|_\infty \leq \delta$, and θ_j and θ_{j+1} differ in one entry only for all $j = 0, \dots n_{\mathcal{A}} - 1$. We conclude that for all $\boldsymbol{x} \in [-1,1]^{d_0}$

$$\|R_\sigma(\theta)(\boldsymbol{x}) - R_\sigma(\theta')(\boldsymbol{x})\|_\infty \leq \sum_{j=0}^{n_{\mathcal{A}}-1} \|R_\sigma(\theta_j)(\boldsymbol{x}) - R_\sigma(\theta_{j+1})(\boldsymbol{x})\|_\infty. \tag{13.1.1}$$

To upper bound (13.1.1), we now only need to understand the effect of changing one weight in a neural network by δ.

Before we can complete the proof we need two auxiliary lemmas. The first of which holds under slightly weaker assumptions of Proposition 13.1.

Lemma 13.2. *Under the assumptions of Proposition 13.1, but with B being allowed to be arbitrary positive, it holds for all $\Phi \in \mathcal{N}(\sigma; \mathcal{A}, B)$*

$$\|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\|_\infty \leq C_\sigma^L \cdot (B d_{\max})^{L+1} \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \tag{13.1.2}$$

for all $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathbb{R}^{d_0}$.

Proof. We start with the case, where $L = 1$. Then, for $(d_0, d_1, d_2) = \mathcal{A}$, we have that

$$\Phi(\boldsymbol{x}) = \mathbf{W}^{(1)} \sigma(\mathbf{W}^{(0)} \boldsymbol{x} + \mathbf{b}^{(0)}) + \mathbf{b}^{(1)},$$

for certain $\mathbf{W}^{(0)}, \mathbf{W}^{(1)}, \mathbf{b}^{(0)}, \mathbf{b}^{(1)}$ with all entries bounded by B. As a consequence, we can estimate

$$\begin{aligned}
\|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\|_\infty &= \left\| \mathbf{W}^{(1)} \left(\sigma(\mathbf{W}^{(0)} \boldsymbol{x} + \mathbf{b}^{(0)}) - \sigma(\mathbf{W}^{(0)} \boldsymbol{x}' + \mathbf{b}^{(0)}) \right) \right\|_\infty \\
&\leq d_1 B \left\| \sigma(\mathbf{W}^{(0)} \boldsymbol{x} + \mathbf{b}^{(0)}) - \sigma(\mathbf{W}^{(0)} \boldsymbol{x}' + \mathbf{b}^{(0)}) \right\|_\infty \\
&\leq d_1 B C_\sigma \left\| \mathbf{W}^{(0)} (\boldsymbol{x} - \boldsymbol{x}') \right\|_\infty \\
&\leq d_1 d_0 B^2 C_\sigma \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_\infty \leq C_\sigma \cdot (d_{\max} B)^2 \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_\infty,
\end{aligned}$$

where we used the Lipschitz property of σ and the fact that $\|\mathbf{A}\boldsymbol{x}\|_\infty \leq n \max_{i,j} |A_{ij}| \|\boldsymbol{x}\|_\infty$ for every matrix $\boldsymbol{A} = (A_{ij})_{i=1,j=1}^{m,n} \in \mathbb{R}^{m \times n}$.

The induction step from L to $L+1$ follows similarly. This concludes the proof of the lemma. □

Lemma 13.3. *Under the assumptions of Proposition 13.1 it holds that*

$$\|\boldsymbol{x}^{(\ell)}\|_\infty \leq (2C_\sigma B d_{\max})^\ell \quad \text{for all } \boldsymbol{x} \in [-1,1]^{d_0}. \tag{13.1.3}$$

Proof. Per Definitions (2.1.1b) and (2.1.1c), we have that for $\ell = 1, \dots, L+1$

$$\begin{aligned}\|\boldsymbol{x}^{(\ell)}\|_\infty &\leq C_\sigma \left\| \mathbf{W}^{(\ell-1)}\boldsymbol{x}^{(\ell-1)} + \mathbf{b}^{(\ell-1)} \right\|_\infty \\ &\leq C_\sigma B d_{\max} \|\boldsymbol{x}^{(\ell-1)}\|_\infty + BC_\sigma,\end{aligned}$$

where we used the triangle inequality and the estimate $\|\mathbf{A}\boldsymbol{x}\|_\infty \leq n \max_{i,j} |A_{ij}| \|\boldsymbol{x}\|_\infty$, which holds for every matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$. We obtain that

$$\begin{aligned}\|\boldsymbol{x}^{(\ell)}\|_\infty &\leq C_\sigma B d_{\max} \cdot (1 + \|\boldsymbol{x}^{(\ell-1)}\|_\infty) \\ &\leq 2C_\sigma B d_{\max} \cdot (\max\{1, \|\boldsymbol{x}^{(\ell-1)}\|_\infty\}).\end{aligned}$$

Resolving the recursive estimate of $\|\boldsymbol{x}^{(\ell)}\|_\infty$ by $2C_\sigma B d_{\max}(\max\{1, \|\boldsymbol{x}^{(\ell-1)}\|_\infty\})$, we conclude that

$$\|\boldsymbol{x}^{(\ell)}\|_\infty \leq (2C_\sigma B d_{\max})^\ell \max\{1, \|\boldsymbol{x}^{(0)}\|_\infty\} = (2C_\sigma B d_{\max})^\ell.$$

This concludes the proof of the lemma. □

We can now proceed with the proof of Proposition 13.1. Assume that θ_{j+1} and θ_j differ only in one entry. We assume this entry to be in the ℓth layer, and we start with the case $\ell < L$. It holds

$$|R_\sigma(\theta_j)(\boldsymbol{x}) - R_\sigma(\theta_{j+1})(\boldsymbol{x})| = |\Phi^\ell(\sigma(\mathbf{W}^{(\ell)}\boldsymbol{x}^{(\ell)} + \mathbf{b}^{(\ell)})) - \Phi^\ell(\sigma(\overline{\mathbf{W}^{(\ell)}}\boldsymbol{x}^{(\ell)} + \overline{\mathbf{b}^{(\ell)}}))|,$$

where $\Phi^\ell \in \mathcal{N}(\sigma; \mathcal{A}^\ell, B)$ for $\mathcal{A}^\ell = (d_{\ell+1}, \dots, d_{L+1})$ and $(\mathbf{W}^{(\ell)}, \mathbf{b}^{(\ell)})$, $(\overline{\mathbf{W}}^{(\ell)}, \overline{\mathbf{b}}^{(\ell)})$ differ in one entry only.

Using the Lipschitz continuity of Φ^ℓ of Lemma 13.2, we have

$$\begin{aligned}&|R_\sigma(\theta_j)(\boldsymbol{x}) - R_\sigma(\theta_{j+1})(\boldsymbol{x})| \\ &\quad\leq C_\sigma^{L-\ell-1}(B d_{\max})^{L-\ell} |\sigma(\mathbf{W}^{(\ell)}\boldsymbol{x}^{(\ell)} + \mathbf{b}^{(\ell)}) - \sigma(\overline{\mathbf{W}^{(\ell)}}\boldsymbol{x}^{(\ell)} + \overline{\mathbf{b}^{(\ell)}})| \\ &\quad\leq C_\sigma^{L-\ell}(B d_{\max})^{L-\ell} \|\mathbf{W}^{(\ell)}\boldsymbol{x}^{(\ell)} + \mathbf{b}^{(\ell)} - \overline{\mathbf{W}^{(\ell)}}\boldsymbol{x}^{(\ell)} - \overline{\mathbf{b}^{(\ell)}}\|_\infty \\ &\quad\leq C_\sigma^{L-\ell}(B d_{\max})^{L-\ell} \delta \max\{1, \|\boldsymbol{x}^{(\ell)}\|_\infty\},\end{aligned}$$

where $\delta \coloneqq \|\theta - \theta'\|_{\max}$. Invoking (13.3), we conclude that

$$\begin{aligned}|R_\sigma(\theta_j)(\boldsymbol{x}) - R_\sigma(\theta_{j+1})(\boldsymbol{x})| &\leq (2C_\sigma B d_{\max})^\ell C_\sigma^{L-\ell} \cdot (B d_{\max})^{L-\ell} \delta \\ &\leq (2C_\sigma B d_{\max})^L \|\theta - \theta'\|_{\max}.\end{aligned}$$

For the case $\ell = L$, a similar estimate can be shown. Combining this with (13.1.1) yields the result. □

Using Proposition 13.1, we can now consider the set of neural networks with a fixed architecture $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ as a subset of $L^\infty([-1,1]^{d_0})$. What is more, is that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is the image of $\mathcal{PN}(\mathcal{A}, \infty)$ under a **locally Lipschitz map**.

13.2 Convexity of neural network spaces

As a first step towards understanding $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ as a subset of $L^\infty([-1,1]^{d_0})$, we notice that it is star-shaped with few centers. Let us first introduce the necessary terminology.

Definition 13.4. Let Z be a subset of a linear space. A point $x \in Z$ is called a **center of** Z if, for every $y \in Z$ it holds that

$$\{tx + (1-t)y \,|\, t \in [0,1]\} \subseteq Z.$$

A set is called **star-shaped** if it has at least one center.

The following proposition follows directly from the definition of a neural network and is the content of Exercise 13.15.

Proposition 13.5. *Let $L \in \mathbb{N}$ and $\mathcal{A} = (d_0, d_1, \dots, d_{L+1}) \in \mathbb{N}^{L+2}$ and $\sigma \colon \mathbb{R} \to \mathbb{R}$. Then $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is scaling invariant, i.e. for every $\lambda \in \mathbb{R}$ it holds that $\lambda f \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$ if $f \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$, and hence $0 \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$ is a center of $\mathcal{N}(\sigma; \mathcal{A}, \infty)$.*

Knowing that $\mathcal{N}(\sigma; \mathcal{A}, B)$ is star-shaped with center 0, we can also ask ourselves if $\mathcal{N}(\sigma; \mathcal{A}, B)$ has more than this one center. It is not hard to see that also every constant function is a center. The following theorem, which corresponds to [171, Proposition C.4], yields an upper bound on the number of *linearly independent* centers.

Theorem 13.6. *Let $L \in \mathbb{N}$ and $\mathcal{A} = (d_0, d_1, \dots, d_{L+1}) \in \mathbb{N}^{L+2}$, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. Then, $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ contains at most $n_\mathcal{A} = \sum_{\ell=0}^{L} (d_\ell + 1) d_{\ell+1}$ linearly independent centers.*

Proof. Assume by contradiction, that there are functions $(g_i)_{i=1}^{n_\mathcal{A}+1} \subseteq \mathcal{N}(\sigma; \mathcal{A}, \infty) \subseteq L^\infty([-1,1]^{d_0})$ that are linearly independent and centers of $\mathcal{N}(\sigma; \mathcal{A}, \infty)$.

By the Theorem of Hahn-Banach, there exist $(g_i')_{i=1}^{n_\mathcal{A}+1} \subseteq (L^\infty([-1,1]^{d_0}))'$ such that $g_i'(g_j) = \delta_{ij}$, for all $i, j \in \{1, \dots, L+1\}$. We define

$$T \colon L^\infty([-1,1]^{d_0}) \to \mathbb{R}^{n_\mathcal{A}+1}, \qquad g \mapsto \begin{pmatrix} g_1'(g) \\ g_2'(g) \\ \vdots \\ g_{n_\mathcal{A}+1}'(g) \end{pmatrix}.$$

Since T is continuous and linear, we have that $T \circ R_\sigma$ is locally Lipschitz continuous by Proposition 13.1. Moreover, since the $(g_i)_{i=1}^{n_\mathcal{A}+1}$ are linearly independent, we have that $T(\operatorname{span}((g_i)_{i=1}^{n_\mathcal{A}+1})) = \mathbb{R}^{n_\mathcal{A}+1}$. We denote $V \coloneqq \operatorname{span}((g_i)_{i=1}^{n_\mathcal{A}+1})$.

Next, we would like to establish that $\mathcal{N}(\sigma;\mathcal{A},\infty) \supset V$. Let $g \in V$ then

$$g = \sum_{\ell=1}^{n_{\mathcal{A}}+1} a_\ell g_\ell,$$

for some $a_1, \dots, a_{n_{\mathcal{A}}+1} \in \mathbb{R}$. We show by induction that $\tilde{g}^{(m)} \coloneqq \sum_{\ell=1}^{m} a_\ell g_\ell \in \mathcal{N}(\sigma;\mathcal{A},\infty)$ for every $m \leq n_{\mathcal{A}}+1$. This is obviously true for $m=1$. Moreover, we have that $\tilde{g}^{(m+1)} = a_{m+1}g_{m+1} + \tilde{g}^{(m)}$. Hence, the induction step holds true if $a_{m+1} = 0$. If $a_{m+1} \neq 0$, then we have that

$$\widetilde{g}^{(m+1)} = 2a_{m+1} \cdot \left(\frac{1}{2} g_{m+1} + \frac{1}{2a_{m+1}} \widetilde{g}^{(m)} \right). \tag{13.2.1}$$

By the induction assumption $\widetilde{g}^{(m)} \in \mathcal{N}(\sigma;\mathcal{A},\infty)$ and hence by Proposition 13.5 $\widetilde{g}^{(m)}/(a_{m+1}) \in \mathcal{N}(\sigma;\mathcal{A},\infty)$. Additionally, since g_{m+1} is a center of $\mathcal{N}(\sigma;\mathcal{A},\infty)$, we have that $\frac{1}{2}g_{m+1} + \frac{1}{2a_{m+1}}\widetilde{g}^{(m)} \in \mathcal{N}(\sigma;\mathcal{A},\infty)$. By Proposition 13.5, we conclude that $\widetilde{g}^{(m+1)} \in \mathcal{N}(\sigma;\mathcal{A},\infty)$.

The induction shows that $g \in \mathcal{N}(\sigma;\mathcal{A},\infty)$ and thus $V \subseteq \mathcal{N}(\sigma;\mathcal{A},\infty)$. As a consequence, $T \circ R_\sigma(\mathcal{PN}(\mathcal{A},\infty)) \supseteq T(V) = \mathbb{R}^{n_{\mathcal{A}}+1}$.

It is a well known fact of basic analysis that for every for $n \in \mathbb{N}$ there does not exist a surjective and locally Lipschitz continuous map from $\mathbb{R}^n$ to $\mathbb{R}^{n+1}$. We recall that $n_{\mathcal{A}} = \dim(\mathcal{PN}(\mathcal{A},\infty))$. This yields the contradiction. □

For a convex set X, the line between all two points of X is a subset of X. Hence, every point of a convex set is a center. This yields the following corollary.

Corollary 13.7. *Let $\mathcal{A} = (d_0, d_1, \dots, d_{L+1})$, let, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. If $\mathcal{N}(\sigma;\mathcal{A},\infty)$ contains more than $n_{\mathcal{A}} = \sum_{\ell=0}^{L}(d_\ell+1)d_{\ell+1}$ linearly independent functions, then $\mathcal{N}(\sigma;\mathcal{A},\infty)$ is not convex.*

Corollary 13.7 tells us that we cannot expect convex sets of neural networks, if the set of neural networks has many linearly independent elements. Sets of neural networks contain for each $f \in \mathcal{N}(\sigma;\mathcal{A},\infty)$ also all shifts of this function, i.e., $f(\cdot + \boldsymbol{b})$ for a $\boldsymbol{b} \in \mathbb{R}^d$ are elements of $f \in \mathcal{N}(\sigma;\mathcal{A},\infty)$. For a set of functions, being shift invariant and having only finitely many linearly independent functions at the same time, is a very restrictive condition. Indeed, it was shown in [171, Proposition C.6] that if $\mathcal{N}(\sigma;\mathcal{A},\infty)$ has only finitely many linearly independent functions and σ is differentiable in at least one point and has non-zero derivative there, then σ is necessarily a polynomial.

We conclude that the set of neural networks is in general non-convex and star-shaped with 0 and constant functions being centers. One could visualize this set in 3D as in Figure 13.1.

The fact, that the neural network space is not convex, could also mean that it merely fails to be convex at one point. For example $\mathbb{R}^2 \setminus \{0\}$ is not convex, but for an optimization algorithm this would likely not pose a problem.

We will next observe that $\mathcal{N}(\sigma;\mathcal{A},\infty)$ does not have such a benign non-convexity and in fact, has *arbitrarily large holes*.

To make this claim mathematically precise, we first introduce the notion of ε-convexity.

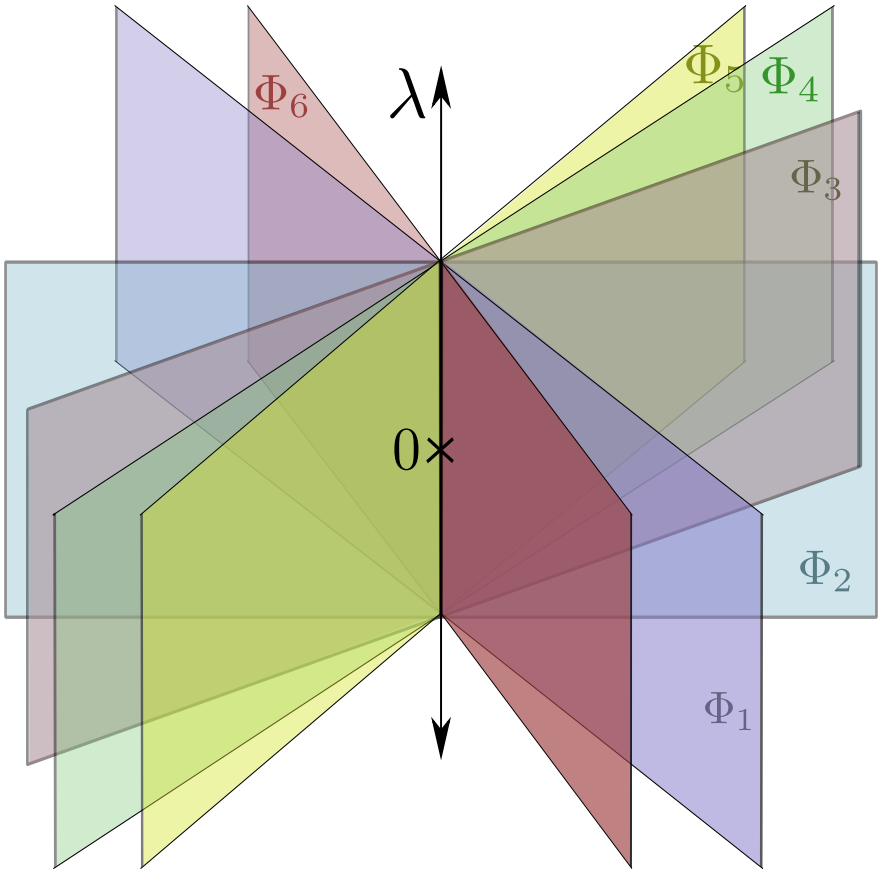

Figure 13.1: Sketch of the space of neural networks in 3D. The vertical axis corresponds to the constant neural network functions, each of which is a center. The set of neural networks consists of many low-dimensional linear subspaces spanned by certain neural networks ($\Phi_1, \ldots, \Phi_6$ in this sketch) and linear functions. Between these low-dimensional subspaces, there is not always a straight-line connection by Corollary 13.7 and Theorem 13.9.

Definition 13.8. For $\varepsilon > 0$, we say that a subset A of a normed vector space X is ε-convex if

$$\mathrm{co}(A) \subseteq A + B_\varepsilon(0),$$

where $\mathrm{co}(A)$ denotes the convex hull of A and $B_\varepsilon(0)$ is an ε ball around 0 with respect to the norm of X.

Intuitively speaking, a set that is convex when one fills up all holes smaller than ε is ε-convex. Now we show that there is no $\varepsilon > 0$ such that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is ε-convex.

Theorem 13.9. *Let $L \in \mathbb{N}$ and $\mathcal{A} = (d_0, d_1, \dots, d_L, 1) \in \mathbb{N}^{L+2}$. Let $K \subseteq \mathbb{R}^{d_0}$ be compact and let $\sigma \in \mathcal{M}$, with $\mathcal{M}$ as in (3.1.1) and assume that σ is not a polynomial. Moreover, assume that there exists an open set, where σ is differentiable and not constant.*

If there exists an $\varepsilon > 0$ such that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is ε-convex, then $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is dense in $C(K)$.

Proof. **Step 1.** We show that ε-convexity implies $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ to be convex. By Proposition 13.5, we have that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is scaling invariant. This implies that $\mathrm{co}(\mathcal{N}(\sigma; \mathcal{A}, \infty))$ is scaling invariant as well. Hence, if there exists $\varepsilon > 0$ such that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is ε-convex, then for every $\varepsilon' > 0$

$$\begin{aligned}\mathrm{co}(\mathcal{N}(\sigma; \mathcal{A}, \infty)) &= \frac{\varepsilon'}{\varepsilon}\mathrm{co}(\mathcal{N}(\sigma; \mathcal{A}, \infty)) \subseteq \frac{\varepsilon'}{\varepsilon}\left(\mathcal{N}(\sigma; \mathcal{A}, \infty) + B_\varepsilon(0)\right) \\ &= \mathcal{N}(\sigma; \mathcal{A}, \infty) + B_{\varepsilon'}(0).\end{aligned}$$

This yields that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is ε'-convex. Since ε' was arbitrary, we have that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is ε-convex for all $\varepsilon > 0$.

As a consequence, we have that

$$\begin{aligned}\mathrm{co}(\mathcal{N}(\sigma; \mathcal{A}, \infty)) &\subseteq \bigcap_{\varepsilon > 0} \left(\mathcal{N}(\sigma; \mathcal{A}, \infty) + B_\varepsilon(0)\right) \\ &\subseteq \bigcap_{\varepsilon > 0} \overline{\left(\mathcal{N}(\sigma; \mathcal{A}, \infty) + B_\varepsilon(0)\right)} = \overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}.\end{aligned}$$

Hence, $\overline{\mathrm{co}}(\mathcal{N}(\sigma; \mathcal{A}, \infty)) \subseteq \overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ and, by the well-known fact that in every metric vector space $\mathrm{co}(\overline{A}) \subseteq \overline{\mathrm{co}}(A)$, we conclude that $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ is convex.

Step 2. We show that $\mathcal{N}_d^1(\sigma; 1) \subseteq \overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$. If $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is ε-convex, then by Step 1 $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ is convex. The scaling invariance of $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ then shows that $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ is a closed linear subspace of $C(K)$.

Note that, by Proposition 3.16 for every $\boldsymbol{w} \in \mathbb{R}^{d_0}$ and $b \in \mathbb{R}$ there exists a function $f \in \overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ such that

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + b) \qquad \text{for all } \boldsymbol{x} \in K. \tag{13.2.2}$$

By definition, every constant function is an element of $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$. Since $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ is a subspace, this implies that all constant functions are in $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$.

Since $\overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$ is a closed vector space, this implies that for all $n \in \mathbb{N}$ and all $\boldsymbol{w}_1^{(1)}, \dots, \boldsymbol{w}_n^{(1)} \in \mathbb{R}^{d_0}$, $w_1^{(2)}, \dots, w_n^{(2)} \in \mathbb{R}$, $b_1^{(1)}, \dots, b_n^{(1)} \in \mathbb{R}$, $b^{(2)} \in \mathbb{R}$

$$\boldsymbol{x} \mapsto \sum_{i=1}^{n} w_i^{(2)} \sigma((\boldsymbol{w}_i^{(1)})^\top \boldsymbol{x} + b_i^{(1)}) + b^{(2)} \in \overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}. \tag{13.2.3}$$

Step 3. From (13.2.3), we conclude that $\mathcal{N}_d^1(\sigma; 1) \subseteq \overline{\mathcal{N}(\sigma; \mathcal{A}, \infty)}$. In words, the whole set of shallow neural networks of arbitrary width is contained in the closure of the set of neural networks with a fixed architecture. By Theorem 3.8, we have that $\mathcal{N}_d^1(\sigma; 1)$ is dense in $C(K)$, which yields the result. □

For any activation function of practical relevance, a set of neural networks with fixed architecture is not dense in $C(K)$. This is only the case for very strange activation functions such as the one discussed in Subsection 3.2. Hence, Theorem 13.9 shows that in general, sets of neural networks of fixed architectures have arbitrarily large holes.

13.3 Closedness and best-approximation property

The non-convexity of the set of neural networks can have some serious consequences for the way we think of the approximation or learning problem by neural networks.

Consider $\mathcal{A} = (d_0, \dots, d_{L+1}) \in \mathbb{N}^{L+2}$ and an activation function σ. Let H be a normed function space on $[-1,1]^{d_0}$ such that $\mathcal{N}(\sigma; \mathcal{A}, \infty) \subseteq H$. For $h \in H$ we would like to find a neural network that best approximates h, i.e. to find $\Phi \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$ such that

$$\|\Phi - h\|_H = \inf_{\Phi^* \in \mathcal{N}(\sigma;\mathcal{A},\infty)} \|\Phi^* - h\|_H. \tag{13.3.1}$$

We say that $\mathcal{N}(\sigma; \mathcal{A}, \infty) \subseteq H$ has

- the **best approximation property**, if for all $h \in H$ there exists at least one $\Phi \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$ such that (13.3.1) holds,
- the **unique best approximation property**, if for all $h \in H$ there exists exactly one $\Phi \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$ such that (13.3.1) holds,
- the **continuous selection property**, if there exists a continuous function $\phi \colon H \to \mathcal{N}(\sigma; \mathcal{A}, \infty)$ such that $\Phi = \phi(h)$ satisfies (13.3.1) for all $h \in H$.

We will see in the sequel, that, in the absence of the best approximation property, we will be able to prove that the learning problem necessarily requires the weights of the neural networks to tend to infinity, which may or may not be desirable in applications.

Moreover, having a continuous selection procedure is desirable as it implies the existence of a stable selection algorithm; that is, an algorithm which, for similar problems yields similar neural networks satisfying (13.3.1).

Below, we will study the properties above for L^p spaces, $p \in [1, \infty)$. As we will see, neural network classes typically neither satisfy the continuous selection nor the best approximation property.

13.3.1 Continuous selection

As shown in [109], neural network spaces essentially never admit the continuous selection property. To give the argument, we first recall the following result from [109, Theorem 3.4] without proof.

Theorem 13.10. *Let $p \in (1, \infty)$. Every subset of $L^p([-1,1]^{d_0})$ with the unique best approximation property is convex.*

This allows to show the next proposition.

Proposition 13.11. *Let $L \in \mathbb{N}$, $\mathcal{A} = (d_0, d_1, \dots, d_{L+1}) \in \mathbb{N}^{L+2}$, let $\sigma : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous and not a polynomial, and let $p \in (1, \infty)$.*

Then, $\mathcal{N}(\sigma; \mathcal{A}, \infty) \subseteq L^p([-1,1]^{d_0})$ does not have the continuous selection property.

Proof. We observe from Theorem 13.6 and the discussion below, that under the assumptions of this result, $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is not convex.

We conclude that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ does not have the unique best approximation property. Moreover, if the set $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ does not have the best approximation property, then it is obvious that it cannot have continuous selection. Thus, we can assume without loss of generality, that $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ has the best approximation property and there exists a point $h \in L^p([-1,1]^{d_0})$ and two different Φ_1,Φ_2 such that

$$\|\Phi_1 - h\|_{L^p} = \|\Phi_2 - h\|_{L^p} = \inf_{\Phi^* \in \mathcal{N}(\sigma; \mathcal{A}, \infty)} \|\Phi^* - h\|_{L^p}. \tag{13.3.2}$$

Note that (13.3.2) implies that $h \notin \mathcal{N}(\sigma; \mathcal{A}, \infty)$.

Let us consider the following function:

$$[-1,1] \ni \lambda \mapsto P(\lambda) = \begin{cases} (1+\lambda)h - \lambda\Phi_1 & \text{for } \lambda \leq 0, \\ (1-\lambda)h + \lambda\Phi_2 & \text{for } \lambda \geq 0. \end{cases}$$

It is clear that $P(\lambda)$ is a continuous path in L^p. Moreover, for $\lambda \in (-1, 0)$

$$\|\Phi_1 - P(\lambda)\|_{L^p} = (1+\lambda)\|\Phi_1 - h\|_{L^p}.$$

Assume towards a contradiction, that there exists $\Phi^* \neq \Phi_1$ such that for $\lambda \in (-1, 0)$

$$\|\Phi^* - P(\lambda)\|_{L^p} \leq \|\Phi_1 - P(\lambda)\|_{L^p}.$$

Then

$$\begin{aligned} \|\Phi^* - h\|_{L^p} &\leq \|\Phi^* - P(\lambda)\|_{L^p} + \|P(\lambda) - h\|_{L^p} \\ &\leq \|\Phi_1 - P(\lambda)\|_{L^p} + \|P(\lambda) - h\|_{L^p} \\ &= (1+\lambda)\|\Phi_1 - h\|_{L^p} + |\lambda|\|\Phi_1 - h\|_{L^p} = \|\Phi_1 - h\|_{L^p}. \end{aligned} \tag{13.3.3}$$

Since Φ_1 is a best approximation to h this implies that every inequality in the estimate above is an equality. Hence, we have that

$$\|\Phi^* - h\|_{L^p} = \|\Phi^* - P(\lambda)\|_{L^p} + \|P(\lambda) - h\|_{L^p}.$$

However, in a strictly convex space like $L^p([-1,1]^{d_0})$ for $p > 1$ this implies that

$$\Phi^* - P(\lambda) = c \cdot (P(\lambda) - h)$$

for a constant $c \neq 0$. This yields that

$$\Phi^* = h + (c+1)\lambda \cdot (h - \Phi_1)$$

and plugging into (13.3.3) yields $|(c+1)\lambda| = 1$. If $(c+1)\lambda = -1$, then we have $\Phi^* = \Phi_1$ which produces a contradiction. If $(c+1)\lambda = 1$, then

$$\begin{aligned}\|\Phi^* - P(\lambda)\|_{L^p} &= \|2h - \Phi_1 - (1+\lambda)h + \lambda\Phi_1\|_{L^p} \\ &= \|(1-\lambda)h - (1-\lambda)\Phi_1\|_{L^p} > \|P(\lambda) - \Phi_1\|_{L^p},\end{aligned}$$

which is another contradiction.

Hence, for every $\lambda < 0$ we have that Φ_1 is the unique minimizer to $P(\lambda)$ in $\mathcal{N}(\sigma; \mathcal{A}, \infty)$. The same argument holds for $\lambda > 0$ and Φ_2. We conclude that for every selection function $\phi\colon L^p([-1,1]^{d_0}) \to \mathcal{N}(\sigma; \mathcal{A}, \infty)$ such that $\Phi = \phi(h)$ satisfies (13.3.1) for all $h \in L^p([-1,1]^{d_0})$ it holds that

$$\lim_{\lambda \downarrow 0} \phi(P(\lambda)) = \Phi_2 \neq \Phi_1 = \lim_{\lambda \uparrow 0} \phi(P(\lambda)).$$

As a consequence, ϕ is not continuous, which shows the result. □

13.3.2 Existence of best approximations

We have seen in Proposition 13.11 that under very mild assumptions, the continuous selection property cannot hold. Moreover, the next result shows that in many cases, also the best approximation property fails to be satisfied. We provide below a simplified version of [171, Theorem 3.1]. We also refer to [68] for earlier work on this problem.

Proposition 13.12. *Let $\mathcal{A} = (1,2,1)$ and let $\sigma : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. Additionally assume that there exist $r > 0$ and $\alpha' \neq \alpha$ such that σ is differentiable for all $|x| > r$ and $\sigma'(x) \to \alpha$ for $x \to \infty$, $\sigma'(x) \to \alpha'$ for $x \to -\infty$.*

Then, there exists a sequence in $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ which converges in $L^p([-1,1]^{d_0})$, for every $p \in (1, \infty)$, and the limit of this sequence is discontinuous. In particular, the limit of the sequence does not lie in $\mathcal{N}(\sigma; \mathcal{A}', \infty)$ for any $\mathcal{A}'$.

Proof. For all $n \in \mathbb{N}$ let

$$f_n(x) = \sigma(nx+1) - \sigma(nx) \qquad \text{for all } x \in \mathbb{R}.$$

Then f_n can be written as a neural network with architecture $(\sigma; 1, 2, 1)$, i.e., $\mathcal{A} = (1,2,1)$. Moreover, for $x > 0$ we observe with the fundamental theorem of calculus and using integration by substitution that

$$f_n(x) = \int_x^{x+1/n} n\sigma'(nz)dz = \int_{nx}^{nx+1} \sigma'(z)dz. \tag{13.3.4}$$

It is not hard to see that the right hand side of (13.3.4) converges to α for $n \to \infty$.

Similarly, for $x < 0$, we observe that $f_n(x)$ converges to α' for $n \to \infty$. We conclude that with $\mathbb{R}_+ = [0, \infty)$ and $\mathbb{R}_- = (-\infty, 0]$

$$f_n \to \alpha \mathbb{1}_{R_+} + \alpha' \mathbb{1}_{\mathbb{R}_-}$$

almost everywhere as $n \to \infty$. Since σ is Lipschitz continuous, we have that f_n is bounded. Therefore, we conclude that $f_n \to \alpha \mathbb{1}_{R_+} + \alpha' \mathbb{1}_{\mathbb{R}_-}$ in L^p for all $p \in [1, \infty)$ by the dominated convergence theorem. □

There is a straight-forward extension of Proposition 13.12 to arbitrary architectures, that will be the content of Exercises 13.16 and 13.17.

Remark 13.13. The proof of Theorem 13.12 does not extend to the L^∞ norm. This, of course, does not mean that generally $\mathcal{N}(\sigma; \mathcal{A}, \infty)$ is a closed set in $L^\infty([-1,1]^{d_0})$. In fact, almost all activation functions used in practice still give rise to non-closed neural network sets, see [171, Theorem 3.3]. However, there is one notable exception. For the ReLU activation function, it can be shown that $\mathcal{N}(\sigma_{\mathrm{ReLU}}; \mathcal{A}, \infty)$ is a closed set in $L^\infty([-1,1]^{d_0})$ if $\mathcal{A}$ has only one hidden layer. The closedness of deep ReLU spaces in L^∞ is an open problem.

13.3.3 Exploding weights phenomenon

Finally, we discuss one of the consequences of the non-existence of best approximations of Proposition 13.12.

Consider a regression problem, where we aim to learn a function f using neural networks with a fixed architecture $\mathcal{N}(\mathcal{A}; \sigma, \infty)$. As discussed in the Chapters 10 and 11, we wish to produce a sequence of neural networks $(\Phi_n)_{n=1}^\infty$ such that the risk defined in (1.2.4) converges to 0. If the loss $\mathcal{L}$ is the squared loss, μ is a probability measure on $[-1,1]^{d_0}$, and the data is given by $(\boldsymbol{x}, f(\boldsymbol{x}))$ for $\boldsymbol{x} \sim \mu$, then

$$\begin{aligned} \mathcal{R}(\Phi_n) &= \|\Phi_n - f\|^2_{L^2([-1,1]^{d_0},\mu)} \\ &= \int_{[-1,1]^{d_0}} |\Phi_n(\boldsymbol{x}) - f(\boldsymbol{x})|^2 d\mu(\boldsymbol{x}) \to 0 \qquad \text{for } n \to \infty. \end{aligned} \tag{13.3.5}$$

According to Proposition 13.12, for a given $\mathcal{A}$, and an activation function σ, it is possible that (13.3.5) holds, but $f \notin \mathcal{N}(\sigma; \mathcal{A}, \infty)$. The following result shows that in this situation, the weights of Φ_n diverge.

Proposition 13.14. *Let $\mathcal{A} = (d_0, d_1, \dots, d_{L+1}) \in \mathbb{N}^{L+2}$, let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous with $C_\sigma \geq 1$, and $|\sigma(x)| \leq C_\sigma |x|$ for all $x \in \mathbb{R}$, and let μ be a measure on $[-1,1]^{d_0}$.*

Assume that there exists a sequence $\Phi_n \in \mathcal{N}(\sigma; \mathcal{A}, \infty)$ and $f \in L^2([-1,1]^{d_0}, \mu) \setminus \mathcal{N}(\sigma; \mathcal{A}, \infty)$ such that

$$\|\Phi_n - f\|^2_{L^2([-1,1]^{d_0},\mu)} \to 0. \tag{13.3.6}$$

Then

$$\limsup_{n\to\infty} \max\left\{ \|\boldsymbol{W}_n^{(\ell)}\|_\infty, \|\boldsymbol{b}_n^{(\ell)}\|_\infty \,\middle|\, \ell = 0, \dots L \right\} = \infty. \tag{13.3.7}$$

Proof. We assume towards a contradiction that the left-hand side of (13.3.7) is finite. As a result, there exists $C > 0$ such that $\Phi_n \in \mathcal{N}(\sigma; \mathcal{A}, C)$ for all $n \in \mathbb{N}$.

By Proposition 13.1, we conclude that $\mathcal{N}(\sigma; \mathcal{A}, C)$ is the image of a compact set under a continuous map and hence is itself a compact set in $L^2([-1,1]^{d_0}, \mu)$. In particular, we have that $\mathcal{N}(\sigma; \mathcal{A}, C)$ is closed. Hence, (13.3.6) implies $f \in \mathcal{N}(\sigma; \mathcal{A}, C)$. This gives a contradiction. □

Proposition 13.14 can be extended to all f for which there is no best approximation in $\mathcal{N}(\sigma; \mathcal{A}, \infty)$, see Exercise 13.18. The results imply that for functions we wish to learn that lack a best approximation within a neural network set, we must expect the weights of the approximating neural networks to grow to infinity. This can be undesirable because, as we will see in the following sections on generalization, a bounded parameter space facilitates many generalization bounds.

Bibliography and further reading

The properties of neural network sets were first studied with a focus on the continuous approximation property in [109, 111, 110] and [68]. The results in [109, 110, 111] already use the non-convexity of sets of shallow neural networks. The results on convexity and closedness presented in this chapter follow mostly the arguments of [171]. Similar results were also derived for other norms [137].

Exercises

Exercise 13.15. Prove Proposition 13.5.

Exercise 13.16. Extend Proposition 13.12 to $\mathcal{A} = (d_0, d_1, 1)$ for arbitrary d_0, $d_1 \in \mathbb{N}$, $d_1 \geq 2$.

Exercise 13.17. Use Proposition 3.16, to extend Proposition 13.12 to arbitrary depth.

Exercise 13.18. Extend Proposition 13.14 to functions f for which there is no best-approximation in $\mathcal{N}(\sigma; \mathcal{A}, \infty)$. To do this, replace (13.3.6) by

$$\|\Phi_n - f\|_{L^2}^2 \to \inf_{\Phi \in \mathcal{N}(\sigma;\mathcal{A},\infty)} \|\Phi - f\|_{L^2}^2.$$