Chapter 3

Universal approximation

After introducing neural networks in Chapter 2, it is natural to inquire about their capabilities. Specifically, we might wonder if there exist inherent limitations to the type of functions a neural network can represent. Could there be a class of functions that neural networks cannot approximate? If so, it would suggest neural networks are specialized tools, similar to how linear regression is suited for linear relationships, but not for data with nonlinear relationships.

In this chapter, we will show that this is not the case, and neural networks are indeed a *universal* tool. More precisely, given sufficiently large and complex architectures, they can approximate almost every sensible input-output relationship. We will formalize and prove this claim in the subsequent sections.

3.1 A universal approximation theorem

To analyze what kind of functions can be approximated with neural networks, we start by considering the uniform approximation of continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ on compact sets. To this end, we first introduce the notion of compact convergence.

Definition 3.1. Let $d \in \mathbb{N}$. A sequence of functions $f_n : \mathbb{R}^d \to \mathbb{R}$, $n \in \mathbb{N}$, is said to **converge compactly** to a function $f : \mathbb{R}^d \to \mathbb{R}$, if for every compact $K \subseteq \mathbb{R}^d$ it holds that $\lim_{n\to\infty} \sup_{\boldsymbol{x}\in K} |f_n(\boldsymbol{x}) - f(\boldsymbol{x})| = 0$. In this case we write $f_n \xrightarrow{\text{cc}} f$.

Throughout what follows, we always consider $C^0(\mathbb{R}^d)$ equipped with the topology of Definition 3.1 (also see Exercise 3.22), and every subset such as $C^0(D)$ with the subspace topology: for example, if $D \subseteq \mathbb{R}^d$ is bounded, then convergence in $C^0(D)$ refers to uniform convergence $\lim_{n\to\infty} \sup_{x\in D} |f_n(x) - f(x)| = 0$.

3.1.1 Universal approximators

As stated before, we want to show that deep neural networks can approximate every continuous function in the sense of Definition 3.1. We call sets of functions that satisfy this property *universal approximators*.

Definition 3.2. Let $d \in \mathbb{N}$. A set of functions $\mathcal{H}$ from $\mathbb{R}^d$ to $\mathbb{R}$ is a **universal approximator** (of $C^0(\mathbb{R}^d)$), if for every $\varepsilon > 0$, every compact $K \subseteq \mathbb{R}^d$, and every $f \in C^0(\mathbb{R}^d)$, there exists $g \in \mathcal{H}$ such that $\sup_{\boldsymbol{x} \in K} |f(\boldsymbol{x}) - g(\boldsymbol{x})| < \varepsilon$.

For a set of (not necessarily continuous) functions $\mathcal{H}$ mapping between $\mathbb{R}^d$ and $\mathbb{R}$, we denote by $\overline{\mathcal{H}}^{\mathrm{cc}}$ its closure with respect to compact convergence.

The relationship between a universal approximator and the closure with respect to compact convergence is established in the proposition below.

Proposition 3.3. *Let $d \in \mathbb{N}$ and $\mathcal{H}$ be a set of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Then, $\mathcal{H}$ is a universal approximator of $C^0(\mathbb{R}^d)$ if and only if $C^0(\mathbb{R}^d) \subseteq \overline{\mathcal{H}}^{\mathrm{cc}}$.*

Proof. Suppose that $\mathcal{H}$ is a universal approximator and fix $f \in C^0(\mathbb{R}^d)$. For $n \in \mathbb{N}$, define $K_n := [-n, n]^d \subseteq \mathbb{R}^d$. Then for every $n \in \mathbb{N}$ there exists $f_n \in \mathcal{H}$ such that $\sup_{\boldsymbol{x} \in K_n} |f_n(\boldsymbol{x}) - f(\boldsymbol{x})| < 1/n$. Since for every compact $K \subseteq \mathbb{R}^d$ there exists n_0 such that $K \subseteq K_n$ for all $n \geq n_0$, it holds $f_n \xrightarrow{\mathrm{cc}} f$. The "only if" part of the assertion is trivial. □

A key tool to show that a set is a universal approximator is the Stone-Weierstrass theorem, see for instance [194, Sec. 5.7].

Theorem 3.4 (Stone-Weierstrass)**.** *Let $d \in \mathbb{N}$, let $K \subseteq \mathbb{R}^d$ be compact, and let $\mathcal{H} \subseteq C^0(K, \mathbb{R})$ satisfy that*

(a) for all $\boldsymbol{x} \in K$ there exists $f \in \mathcal{H}$ such that $f(\boldsymbol{x}) \neq 0$,

(b) for all $\boldsymbol{x} \neq \boldsymbol{y} \in K$ there exists $f \in \mathcal{H}$ such that $f(\boldsymbol{x}) \neq f(\boldsymbol{y})$,

(c) $\mathcal{H}$ is an algebra of functions, i.e., $\mathcal{H}$ is closed under addition, multiplication and scalar multiplication.

Then $\mathcal{H}$ is dense in $C^0(K)$.

Example 3.5 (Polynomials are dense in $C^0(\mathbb{R}^d)$)**.** For a multiindex $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ and a vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ denote $\boldsymbol{x}^{\boldsymbol{\alpha}} := \prod_{j=1}^d x_j^{\alpha_j}$. In the following, with $|\boldsymbol{\alpha}| := \sum_{j=1}^d \alpha_j$, we write

$$\mathbb{P}_n := \operatorname{span}\{\boldsymbol{x}^{\boldsymbol{\alpha}} \,|\, \boldsymbol{\alpha} \in \mathbb{N}_0^d,\ |\boldsymbol{\alpha}| \leq n\}$$

i.e., $\mathbb{P}_n$ is the space of polynomials of degree at most n (with real coefficients). It is easy to check that $\mathbb{P} := \bigcup_{n \in \mathbb{N}} \mathbb{P}_n(\mathbb{R}^d)$ satisfies the assumptions of Theorem 3.4 on every compact set $K \subseteq \mathbb{R}^d$. Thus the space of polynomials $\mathbb{P}$ is a universal approximator of $C^0(\mathbb{R}^d)$, and by Proposition 3.3, $\mathbb{P}$ is dense in $C^0(\mathbb{R}^d)$. In case we wish to emphasize the dimension of the underlying space, in the following we will also write $\mathbb{P}_n(\mathbb{R}^d)$ or $\mathbb{P}(\mathbb{R}^d)$ to denote $\mathbb{P}_n$, $\mathbb{P}$ respectively.

3.1.2 Shallow neural networks

With the necessary formalism established in the previous subsection, we can now demonstrate that shallow neural networks of arbitrary width form a universal approximator under certain (mild) conditions on the activation function. The results in this section are based on [130], and for the proofs we follow the arguments in that paper.

We first introduce notation for the set of all functions realized by certain architectures.

Definition 3.6. Let d, m, L, $n \in \mathbb{N}$ and $\sigma\colon \mathbb{R} \to \mathbb{R}$. The set of all functions realized by neural networks with d-dimensional input, m-dimensional output, depth at most L, width at most n, and activation function σ is denoted by

$$\mathcal{N}_d^m(\sigma; L, n) := \{\Phi : \mathbb{R}^d \to \mathbb{R}^m \,|\, \Phi \text{ as in Def. 2.1, } \mathrm{depth}(\Phi) \leq L, \ \mathrm{width}(\Phi) \leq n\}.$$

Furthermore,

$$\mathcal{N}_d^m(\sigma; L) := \bigcup_{n \in \mathbb{N}} \mathcal{N}_d^m(\sigma; L, n).$$

In the sequel, we require the activation function σ to belong to the set of piecewise continuous and locally bounded functions

$$\begin{aligned} \mathcal{M} := \big\{ \sigma \in L^\infty_{\mathrm{loc}}(\mathbb{R}) \;\big|\; &\text{there exist intervals } I_1, \dots, I_M \text{ partitioning } \mathbb{R}, \\ &\text{s.t. } \sigma \in C^0(I_j) \text{ for all } j = 1, \dots, M \big\}. \end{aligned} \tag{3.1.1}$$

Here, $M \in \mathbb{N}$ is finite, and the intervals I_j are understood to have positive (possibly infinite) Lebesgue measure, i.e. I_j is e.g. not allowed to be empty or a single point. Hence, σ is a piecewise continuous function, and it has discontinuities at most finitely many points.

Example 3.7. Activation functions belonging to $\mathcal{M}$ include, in particular, all continuous non-polynomial functions, which in turn includes all practically relevant activation functions such as the ReLU, the SiLU, and the Sigmoid discussed in Section 2.3. In these cases, we can choose $M = 1$ and $I_1 = \mathbb{R}$. Discontinuous functions include for example the Heaviside function $x \mapsto \mathbb{1}_{x>0}$ (also called a "perceptron" in this context) but also $x \mapsto \mathbb{1}_{x>0} \sin(1/x)$: Both belong to $\mathcal{M}$ with $M = 2$, $I_1 = (-\infty, 0]$ and $I_2 = (0, \infty)$. We exclude for example the function $x \mapsto 1/x$, which is not locally bounded.

The rest of this subsection is dedicated to proving the following theorem that has now already been announced repeatedly.

Theorem 3.8. *Let $d \in \mathbb{N}$ and $\sigma \in \mathcal{M}$. Then $\mathcal{N}_d^1(\sigma; 1)$ is a universal approximator of $C^0(\mathbb{R}^d)$ if and only if σ is not a polynomial.*

Remark 3.9. We will see in Exercise 3.26 and Corollary 3.18 that neural networks can also arbitrarily well approximate non-continuous functions with respect to suitable norms.

The universal approximation theorem by Leshno, Lin, Pinkus and Schocken [130]—of which Theorem 3.8 is a special case—is even formulated for a much larger set $\mathcal{M}$, which allows for activation functions that have discontinuities at a (possibly non-finite) set of Lebesgue measure zero. Instead of proving the theorem in this generality, we resort to the simpler case stated above. This allows to avoid some technicalities, but the main ideas remain the same. The proof strategy is to verify the following three claims:

(i) if $C^0(\mathbb{R}^1) \subseteq \overline{\mathcal{N}_1^1(\sigma;1)}^{\mathrm{cc}}$ then $C^0(\mathbb{R}^d) \subseteq \overline{\mathcal{N}_d^1(\sigma;1)}^{\mathrm{cc}}$,

(ii) if $\sigma \in C^\infty(\mathbb{R})$ is not a polynomial then $C^0(\mathbb{R}^1) \subseteq \overline{\mathcal{N}_1^1(\sigma;1)}^{\mathrm{cc}}$,

(iii) if $\sigma \in \mathcal{M}$ is not a polynomial then there exists $\tilde{\sigma} \in C^\infty(\mathbb{R}) \cap \overline{\mathcal{N}_1^1(\sigma;1)}^{\mathrm{cc}}$ which is not a polynomial.

Upon observing that $\tilde{\sigma} \in \overline{\mathcal{N}_1^1(\sigma;1)}^{\mathrm{cc}}$ implies $\overline{\mathcal{N}_1^1(\tilde{\sigma},1)}^{\mathrm{cc}} \subseteq \overline{\mathcal{N}_1^1(\sigma;1)}^{\mathrm{cc}}$, it is easy to see that these statements together with Proposition 3.3 establish the implication "$\Leftarrow$" asserted in Theorem 3.8. The reverse direction is straightforward to check and will be the content of Exercise 3.23.

We start with a more general version of (i) and reduce the problem to the one dimensional case.

Lemma 3.10. *Assume that $\mathcal{H}$ is a universal approximator of $C^0(\mathbb{R})$. Then for every $d \in \mathbb{N}$*

$$\operatorname{span}\{\boldsymbol{x} \mapsto g(\boldsymbol{w}\cdot\boldsymbol{x}) \,|\, \boldsymbol{w} \in \mathbb{R}^d,\ g \in \mathcal{H}\}$$

is a universal approximator of $C^0(\mathbb{R}^d)$.

Proof. For $k \in \mathbb{N}_0$, denote by $\mathbb{H}_k$ the space of all k-homogenous polynomials, that is

$$\mathbb{H}_k := \operatorname{span}\left\{\mathbb{R}^d \ni \boldsymbol{x} \mapsto \boldsymbol{x}^{\boldsymbol{\alpha}} \,\middle|\, \boldsymbol{\alpha} \in \mathbb{N}_0^d,\ |\boldsymbol{\alpha}| = k\right\}.$$

We claim that

$$\mathbb{H}_k \subseteq \overline{\operatorname{span}\{\mathbb{R}^d \ni \boldsymbol{x} \mapsto g(\boldsymbol{w}\cdot\boldsymbol{x}) \,|\, \boldsymbol{w} \in \mathbb{R}^d,\ g \in \mathcal{H}\}}^{\mathrm{cc}} =: X \tag{3.1.2}$$

for all $k \in \mathbb{N}_0$. This implies that all multivariate polynomials belong to X. An application of the Stone-Weierstrass theorem (cp. Example 3.5) and Proposition 3.3 then conclude the proof.

For every $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = k$, it holds $D^{\boldsymbol{\beta}}\boldsymbol{x}^{\boldsymbol{\alpha}} = \delta_{\boldsymbol{\beta},\boldsymbol{\alpha}}\boldsymbol{\alpha}!$, where $\boldsymbol{\alpha}! := \prod_{j=1}^d \alpha_j!$ and $\delta_{\boldsymbol{\beta},\boldsymbol{\alpha}} = 1$ if $\boldsymbol{\beta} = \boldsymbol{\alpha}$ and $\delta_{\boldsymbol{\beta},\boldsymbol{\alpha}} = 0$ otherwise. Hence, since $\{\boldsymbol{x} \mapsto \boldsymbol{x}^{\boldsymbol{\alpha}} \,|\, |\boldsymbol{\alpha}| = k\}$ is a basis of $\mathbb{H}_k$, the set $\{D^{\boldsymbol{\alpha}} \,|\, |\boldsymbol{\alpha}| = k\}$ is a basis of its topological dual $\mathbb{H}_k'$. Thus each linear functional $l \in \mathbb{H}_k'$ allows the representation $l = p(D)$ for some $p \in \mathbb{H}_k$ (here D stands for the differential).

By the multinomial formula

$$(\boldsymbol{w}\cdot\boldsymbol{x})^k = \left(\sum_{j=1}^d w_j x_j\right)^k = \sum_{\{\boldsymbol{\alpha}\in\mathbb{N}_0^d \,|\, |\boldsymbol{\alpha}|=k\}} \frac{k!}{\boldsymbol{\alpha}!}\boldsymbol{w}^{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}}.$$

Therefore, we have that $(\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k) \in \mathbb{H}_k$. Moreover, for every $l = p(D) \in \mathbb{H}_k'$ and all $\boldsymbol{w} \in \mathbb{R}^d$ we have that

$$l(\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k) = k! p(\boldsymbol{w}).$$

Hence, if $l(\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k) = p(D)(\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k) = 0$ for all $\boldsymbol{w} \in \mathbb{R}^d$, then $p \equiv 0$ and thus $l \equiv 0$.

This implies $\operatorname{span}\{\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k \,|\, \boldsymbol{w} \in \mathbb{R}^d\} = \mathbb{H}_k$. Indeed, if there exists $h \in \mathbb{H}_k$ which is not in $\operatorname{span}\{\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k \,|\, \boldsymbol{w} \in \mathbb{R}^d\}$, then by the theorem of Hahn-Banach (see Theorem B.8), there exists a non-zero functional in $\mathbb{H}_k'$ vanishing on $\operatorname{span}\{\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k \,|\, \boldsymbol{w} \in \mathbb{R}^d\}$. This contradicts the previous observation.

By the universality of $\mathcal{H}$ it is not hard to see that $\boldsymbol{x} \mapsto (\boldsymbol{w} \cdot \boldsymbol{x})^k \in X$ for all $\boldsymbol{w} \in \mathbb{R}^d$. Therefore, we have $\mathbb{H}_k \subseteq X$ for all $k \in \mathbb{N}_0$. □

By the above lemma, in order to verify that $\mathcal{N}_d^1(\sigma; 1)$ is a universal approximator, it suffices to show that $\mathcal{N}_1^1(\sigma; 1)$ is a universal approximator. We first show that this is the case for sigmoidal activations.

Definition 3.11. An activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is called **sigmoidal**, if $\sigma \in C^0(\mathbb{R})$, $\lim_{x \to \infty} \sigma(x) = 1$ and $\lim_{x \to -\infty} \sigma(x) = 0$.

For sigmoidal activation functions we can now conclude the universality in the univariate case.

Lemma 3.12. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be monotonically increasing and sigmoidal. Then $C^0(\mathbb{R}) \subseteq \overline{\mathcal{N}_1^1(\sigma; 1)}^{\mathrm{cc}}$.*

We prove Lemma 3.12 in Exercise 3.24. Lemma 3.10 and Lemma 3.12 show Theorem 3.8 in the special case where σ is monotonically increasing and sigmoidal. For the general case, let us continue with (ii) and consider C^∞ activations.

Lemma 3.13. *If $\sigma \in \mathbb{C}^\infty(\mathbb{R})$ and σ is not a polynomial, then $\mathcal{N}_1^1(\sigma; 1)$ is dense in $C^0(\mathbb{R})$.*

Proof. Denote $X := \overline{\mathcal{N}_1^1(\sigma; 1)}^{\mathrm{cc}}$. We show again that all polynomials belong to X. An application of the Stone-Weierstrass theorem then gives the statement.

Fix $b \in \mathbb{R}$ and denote $f_x(w) := \sigma(wx + b)$ for all x, $w \in \mathbb{R}$. By Taylor's theorem, for $h \neq 0$

$$\begin{aligned} \frac{\sigma((w+h)x + b) - \sigma(wx + b)}{h} &= \frac{f_x(w+h) - f_x(w)}{h} \\ &= f_x'(w) + \frac{h}{2} f_x''(\xi) \\ &= f_x'(w) + \frac{h}{2} x^2 \sigma''(\xi x + b) \end{aligned} \tag{3.1.3}$$

for some $\xi = \xi(h)$ between w and $w+h$. Note that the left-hand side belongs to $\mathcal{N}_1^1(\sigma;1)$ as a function of x. Since $\sigma'' \in C^0(\mathbb{R})$, for every compact set $K \subseteq \mathbb{R}$

$$\sup_{x\in K}\sup_{|h|\leq 1}|x^2\sigma''(\xi(h)x+b)| \leq \sup_{x\in K}\sup_{\eta\in[w-1,w+1]}|x^2\sigma''(\eta x+b)| < \infty.$$

Letting $h \to 0$, as a function of x the term in (3.1.3) thus converges uniformly towards $K \ni x \mapsto f_x'(w)$. Since K was arbitrary, $x \mapsto f_x'(w)$ belongs to X. Inductively applying the same argument to $f_x^{(k-1)}(w)$, we find that $x \mapsto f_x^{(k)}(w)$ belongs to X for all $k \in \mathbb{N}$, $w \in \mathbb{R}$. Observe that $f_x^{(k)}(w) = x^k\sigma^{(k)}(wx+b)$. Since σ is not a polynomial, for each $k \in \mathbb{N}$ there exists $b_k \in \mathbb{R}$ such that $\sigma^{(k)}(b_k) \neq 0$. Choosing $w = 0$, we obtain that $x \mapsto x^k$ belongs to X. □

Finally, we come to the proof of (iii)—the claim that there exists at least one non-polynomial $C^\infty(\mathbb{R})$ function in the closure of $\mathcal{N}_1^1(\sigma;1)$. The argument is split into two lemmata. Denote in the following by $C_c^\infty(\mathbb{R})$ the set of compactly supported $C^\infty(\mathbb{R})$ functions.

Lemma 3.14. *Let $\sigma \in \mathcal{M}$. Then for each $\varphi \in C_c^\infty(\mathbb{R})$ it holds $\sigma * \varphi \in \overline{\mathcal{N}_1^1(\sigma;1)}^{\mathrm{cc}}$.*

Proof. Fix $\varphi \in C_c^\infty(\mathbb{R})$ and let $a > 0$ such that $\operatorname{supp}\varphi \subseteq [-a,a]$. We have

$$\sigma * \varphi(x) = \int_{\mathbb{R}} \sigma(x-y)\varphi(y)\,\mathrm{d}y.$$

Denote $y_j := -a + 2aj/n$ for $j = 0,\dots,n$ and define for $x \in \mathbb{R}$

$$f_n(x) := \frac{2a}{n}\sum_{j=0}^{n-1}\sigma(x-y_j)\varphi(y_j).$$

Clearly, $f_n \in \mathcal{N}_1^1(\sigma;1)$. We will show $f_n \xrightarrow{\mathrm{cc}} \sigma*\varphi$ as $n \to \infty$. To do so we verify uniform convergence of f_n towards $\sigma * \varphi$ on the interval $[-b,b]$ with $b > 0$ arbitrary but fixed.

For $x \in [-b,b]$

$$|\sigma*\varphi(x) - f_n(x)| \leq \sum_{j=0}^{n-1}\left|\int_{y_j}^{y_{j+1}}\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)\,\mathrm{d}y\right|. \tag{3.1.4}$$

Fix $\varepsilon \in (0,1)$. Since $\sigma \in \mathcal{M}$, there exist $z_1,\dots,z_M \in \mathbb{R}$ such that σ is continuous on $\mathbb{R}\backslash\{z_1,\dots,z_M\}$ (cp. (3.1.1)). With $D_\varepsilon := \bigcup_{j=1}^M (z_j-\varepsilon, z_j+\varepsilon)$, observe that σ is uniformly continuous on the compact set $K_\varepsilon := [-a-b, a+b] \cap D_\varepsilon^c$. Now let $J_c \cup J_d = \{0,\dots,n-1\}$ be a partition (depending on x), such that $j \in J_c$ if and only if $[x-y_{j+1}, x-y_j] \subseteq K_\varepsilon$. Hence, $j \in J_d$ implies the existence of $i \in \{1,\dots,M\}$ such that the distance of z_i to $[x-y_{j+1}, x-y_j]$ is at most ε. Due to the interval

$[x - y_{j+1}, x - y_j]$ having length $2a/n$, we can bound

$$\begin{aligned}\sum_{j\in J_d} y_{j+1} - y_j &= \left|\bigcup_{j\in J_d} [x - y_{j+1}, x - y_j]\right| \\ &\le \left|\bigcup_{i=1}^{M} \Big[z_i - \varepsilon - \frac{2a}{n}, z_i + \varepsilon + \frac{2a}{n}\Big]\right| \\ &\le M \cdot \Big(2\varepsilon + \frac{4a}{n}\Big).\end{aligned}$$

Next, because of the local boundedness of σ and the fact that $\varphi \in C_c^\infty$, it holds $\sup_{|y|\le a+b} |\sigma(y)| + \sup_{|y|\le a} |\varphi(y)| =: \gamma < \infty$. Hence

$$\begin{aligned}&|\sigma * \varphi(x) - f_n(x)| \\ &\quad\le \sum_{j\in J_c\cup J_d} \left|\int_{y_j}^{y_{j+1}} \sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)\,\mathrm{d}y\right| \\ &\quad\le 2\gamma^2 M \cdot \left(2\varepsilon + \frac{4a}{n}\right) \\ &\qquad + 2a \sup_{j\in J_c} \max_{y\in[y_j, y_{j+1}]} |\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)|. \qquad (3.1.5)\end{aligned}$$

We can bound the term in the last maximum by

$$\begin{aligned}&|\sigma(x-y)\varphi(y) - \sigma(x-y_j)\varphi(y_j)| \\ &\quad\le |\sigma(x-y) - \sigma(x-y_j)||\varphi(y)| + |\sigma(x-y_j)||\varphi(y) - \varphi(y_j)| \\ &\quad\le \gamma \cdot \left(\sup_{\substack{z_1,z_2\in K_\varepsilon \\ |z_1-z_2|\le \frac{2a}{n}}} |\sigma(z_1) - \sigma(z_2)| + \sup_{\substack{z_1,z_2\in[-a,a] \\ |z_1-z_2|\le \frac{2a}{n}}} |\varphi(z_1) - \varphi(z_2)|\right).\end{aligned}$$

Finally, uniform continuity of σ on K_ε and φ on $[-a, a]$ imply that the last term tends to 0 as $n \to \infty$ uniformly for all $x \in [-b, b]$. This shows that there exist $C < \infty$ (independent of ε and x) and $n_\varepsilon \in \mathbb{N}$ (independent of x) such that the term in (3.1.5) is bounded by $C\varepsilon$ for all $n \ge n_\varepsilon$. Since ε was arbitrary, this yields the claim. □

Lemma 3.15. *If $\sigma \in \mathcal{M}$ and $\sigma * \varphi$ is a polynomial for all $\varphi \in C_c^\infty(\mathbb{R})$, then σ is a polynomial.*

Proof. Fix $-\infty < a < b < \infty$ and consider $C_c^\infty(a, b) := \{\varphi \in C^\infty(\mathbb{R}) \mid \operatorname{supp}\varphi \subseteq [a, b]\}$. Define a metric ρ on $C_c^\infty(a, b)$ via

$$\rho(\varphi, \psi) := \sum_{j\in\mathbb{N}_0} 2^{-j} \frac{|\varphi - \psi|_{C^j(a,b)}}{1 + |\varphi - \psi|_{C^j(a,b)}},$$

where

$$|\varphi|_{C^j(a,b)} := \sup_{x\in[a,b]} |\varphi^{(j)}(x)|.$$

Since the space of j times differentiable functions on $[a,b]$ is complete with respect to the norm $\sum_{i=0}^{j} |\cdot|_{C^i(a,b)}$, see for instance [88, Satz 104.3], the space $C_c^\infty(a,b)$ is complete with the metric ρ. For $k \in \mathbb{N}$ set

$$V_k := \{\varphi \in C_c^\infty(a,b) \,|\, \sigma * \varphi \in \mathbb{P}_k\},$$

where $\mathbb{P}_k := \operatorname{span}\{\mathbb{R} \ni x \mapsto x^j \,|\, 0 \le j \le k\}$ denotes the space of polynomials of degree at most k. Then V_k is closed with respect to the metric ρ. To see this, we only need to observe that for a converging sequence $\varphi_j \to \varphi^*$ with respect to ρ and $\varphi_j \in V_k$, it follows that $D^{k+1}(\sigma * \varphi^*) = 0$ and hence $\sigma * \varphi^*$ is a polynomial. Since $D^{k+1}(\sigma * \varphi_j) = 0$ we compute with the linearity of the convolution and the fact that $D^{k+1}(f * g) = f * D^{k+1}(g)$ for differentiable g and if both sides are well-defined that

$$\begin{aligned}
&\sup_{x\in[a,b]} |D^{k+1}(\sigma * \varphi^*)(x)| \\
&\quad = \sup_{x\in[a,b]} |\sigma * D^{k+1}(\varphi^* - \varphi_j)(x)| \\
&\quad \le |b-a| \sup_{z\in[a-b,b-a]} |\sigma(z)| \cdot \sup_{x\in[a,b]} |D^{k+1}(\varphi_j - \varphi^*)(x)|
\end{aligned}$$

and since σ is locally bounded, the right hand-side converges to 0.

By assumption we have

$$\bigcup_{k\in\mathbb{N}} V_k = C_c^\infty(a,b).$$

Baire's category theorem implies the existence of $k_0 \in \mathbb{N}$ (depending on a, b) such that V_{k_0} contains an open subset of $C_c^\infty(a,b)$. Since V_{k_0} is a vector space, it must hold $V_{k_0} = C_c^\infty(a,b)$.

We now show that $\varphi * \sigma \in \mathbb{P}_{k_0}$ for every $\varphi \in C_c^\infty(\mathbb{R})$; in other words, $k_0 = k_0(a,b)$ can be chosen independent of a and b. First consider a shift $s \in \mathbb{R}$ and let $\tilde{a} := a + s$ and $\tilde{b} := b + s$. Then with $S(x) := x + s$, for any $\varphi \in C_c^\infty(\tilde{a},\tilde{b})$ holds $\varphi \circ S \in C_c^\infty(a,b)$, and thus $(\varphi \circ S) * \sigma \in \mathbb{P}_{k_0}$. Since $(\varphi \circ S) * \sigma(x) = \varphi * \sigma(x+s)$, we conclude that $\varphi * \sigma \in \mathbb{P}_{k_0}$. Next let $-\infty < \tilde{a} < \tilde{b} < \infty$ be *arbitrary*. Then, for an integer $n > (\tilde{b} - \tilde{a})(b-a)$ we can cover $(\tilde{a},\tilde{b})$ with $n \in \mathbb{N}$ overlapping open intervals $(a_1,b_1),\dots,(a_n,b_n)$, each of length $b-a$. Any $\varphi \in C_c^\infty(\tilde{a},\tilde{b})$ can be written as $\varphi = \sum_{j=1}^n \varphi_j$ where $\varphi_j \in C_c^\infty(a_j,b_j)$. Then $\varphi * \sigma = \sum_{j=1}^n \varphi_j * \sigma \in \mathbb{P}_{k_0}$, and thus $\varphi * \sigma \in \mathbb{P}_{k_0}$ for every $\varphi \in C_c^\infty(\mathbb{R})$.

Finally, Exercise 3.25 implies $\sigma \in \mathbb{P}_{k_0}$. □

Now we can put everything together to show Theorem 3.8.

of Theorem 3.8. By Exercise 3.23 we have the implication "⇒".

For the other direction we assume that $\sigma \in \mathcal{M}$ is not a polynomial. Then by Lemma 3.15 there exists $\varphi \in C_c^\infty(\mathbb{R})$ such that $\sigma * \varphi$ is not a polynomial. According to Lemma 3.14 we have $\sigma * \varphi \in \overline{\mathcal{N}_1^1(\sigma;1)}^{cc}$. We conclude with Lemma 3.13 that $\mathcal{N}_1^1(\sigma;1)$ is a universal approximator of $C^0(\mathbb{R})$.

Finally, by Lemma 3.10, $\mathcal{N}_d^1(\sigma;1)$ is a universal approximator of $C^0(\mathbb{R}^d)$. □

3.1.3 Deep neural networks

Theorem 3.8 shows the universal approximation capability of single-hidden-layer neural networks with activation functions $\sigma \in \mathcal{M}\backslash\mathbb{P}$: they can approximate every continuous function on every compact set to arbitrary precision, given sufficient width. This result directly extends to neural networks of any fixed depth $L \geq 1$. The idea is to use the fact that the identity function can be approximated with a shallow neural network. By composing a shallow neural network approximation of the target function f with (multiple) shallow neural networks approximating the identity function, gives a deep neural network approximation of f.

Instead of directly applying Theorem 3.8, we first establish the following proposition regarding the approximation of the identity function. Rather than $\sigma \in \mathcal{M}\backslash\mathbb{P}$, it requires a different (mild) assumption on the activation function. This allows for a constructive proof, yielding explicit bounds on the neural network size, which will prove useful later in the book.

Proposition 3.16. *Let d, $L \in \mathbb{N}$, let $K \subseteq \mathbb{R}^d$ be compact, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be such that there exists an open set on which σ is differentiable and not constant. Then, for every $\varepsilon > 0$, there exists a neural network $\Phi \in \mathcal{N}_d^d(\sigma; L, d)$ such that*

$$\|\Phi(\boldsymbol{x}) - \boldsymbol{x}\|_\infty < \varepsilon \qquad \text{for all } \boldsymbol{x} \in K.$$

Proof. The proof uses the same idea as in Lemma 3.13, where we approximate the derivative of the activation function by a simple neural network. Let us first assume $d \in \mathbb{N}$ and $L = 1$.

Let $x^* \in \mathbb{R}$ be such that σ is differentiable on a neighborhood of x^* and $\sigma'(x^*) = \theta \neq 0$. Moreover, let $\boldsymbol{x}^* = (x^*, \dots, x^*) \in \mathbb{R}^d$. Then, for $\lambda > 0$ we define

$$\Phi_\lambda(\boldsymbol{x}) \coloneqq \frac{\lambda}{\theta}\sigma\left(\frac{\boldsymbol{x}}{\lambda} + \boldsymbol{x}^*\right) - \frac{\lambda}{\theta}\sigma(\boldsymbol{x}^*),$$

Then, we have, for all $\boldsymbol{x} \in K$,

$$\Phi_\lambda(\boldsymbol{x}) - \boldsymbol{x} = \lambda\frac{\sigma(\boldsymbol{x}/\lambda + \boldsymbol{x}^*) - \sigma(\boldsymbol{x}^*)}{\theta} - \boldsymbol{x}. \tag{3.1.6}$$

If $x_i = 0$ for $i \in \{1, \dots, d\}$, then (3.1.6) shows that $(\Phi_\lambda(\boldsymbol{x}) - \boldsymbol{x})_i = 0$. Otherwise

$$|(\Phi_\lambda(\boldsymbol{x}) - \boldsymbol{x})_i| = \frac{|x_i|}{|\theta|}\left|\frac{\sigma(x_i/\lambda + x^*) - \sigma(x^*)}{x_i/\lambda} - \theta\right|.$$

By the definition of the derivative, we have that $|(\Phi_\lambda(\boldsymbol{x}) - \boldsymbol{x})_i| \to 0$ for $\lambda \to \infty$ uniformly for all $\boldsymbol{x} \in K$ and $i \in \{1, \dots, d\}$. Therefore, $|\Phi_\lambda(\boldsymbol{x}) - \boldsymbol{x}| \to 0$ for $\lambda \to \infty$ uniformly for all $\boldsymbol{x} \in K$.

The extension to $L > 1$ is straight forward and is the content of Exercise 3.27. □

Using the aforementioned generalization of Proposition 3.16 to arbitrary non-polynomial activation functions $\sigma \in \mathcal{M}$, we obtain the following extension of Theorem 3.8.

Corollary 3.17. *Let $d \in \mathbb{N}$, $L \in \mathbb{N}$ and $\sigma \in \mathcal{M}$. Then $\mathcal{N}_d^1(\sigma; L)$ is a universal approximator of $C^0(\mathbb{R}^d)$ if and only if σ is not a polynomial.*

Proof. We only show the implication "$\Leftarrow$". The other direction is again left as an exercise, see Exercise 3.23.

Assume $\sigma \in \mathcal{M}$ is not a polynomial, let $K \subseteq \mathbb{R}^d$ be compact, and let $f \in C^0(\mathbb{R}^d)$. Fix $\varepsilon \in (0, 1)$. We need to show that there exists a neural network $\Phi \in \mathcal{N}_d^1(\sigma; L)$ such that $\sup_{\boldsymbol{x} \in K} |f(\boldsymbol{x}) - \Phi(\boldsymbol{x})| < \varepsilon$. The case $L = 1$ holds by Theorem 3.8, so let $L > 1$.

By Theorem 3.8, there exist $\Phi_{\text{shallow}} \in \mathcal{N}_d^1(\sigma; 1)$ such that

$$\sup_{\boldsymbol{x} \in K} |f(\boldsymbol{x}) - \Phi_{\text{shallow}}(\boldsymbol{x})| < \frac{\varepsilon}{2}. \tag{3.1.7}$$

Compactness of $\{f(\boldsymbol{x}) \,|\, \boldsymbol{x} \in K\}$ implies that we can find $n > 0$ such that

$$\{\Phi_{\text{shallow}}(\boldsymbol{x}) \,|\, \boldsymbol{x} \in K\} \subseteq [-n, n]. \tag{3.1.8}$$

Let $\Phi_{\text{id}} \in \mathcal{N}_1^1(\sigma; L - 1)$ be an approximation to the identity such that

$$\sup_{x \in [-n,n]} |x - \Phi_{\text{id}}(x)| < \frac{\varepsilon}{2}, \tag{3.1.9}$$

which is possibly by the extension of Proposition 3.16 to general non-polynomial activation functions $\sigma \in \mathcal{M}$.

Denote $\Phi := \Phi_{\text{id}} \circ \Phi_{\text{shallow}}$. According to Proposition 2.3 (iv) holds $\Phi \in \mathcal{N}_d^1(\sigma; L)$ as desired. Moreover (3.1.7), (3.1.8), (3.1.9) imply

$$\begin{aligned}
\sup_{\boldsymbol{x} \in K} |f(\boldsymbol{x}) - \Phi(\boldsymbol{x})| &= \sup_{\boldsymbol{x} \in K} |f(\boldsymbol{x}) - \Phi_{\text{id}}(\Phi_{\text{shallow}}(\boldsymbol{x}))| \\
&\leq \sup_{\boldsymbol{x} \in K} \big(|f(\boldsymbol{x}) - \Phi_{\text{shallow}}(\boldsymbol{x})| + |\Phi_{\text{shallow}}(\boldsymbol{x}) - \Phi_{\text{id}}(\Phi_{\text{shallow}}(\boldsymbol{x}))| \big) \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

This concludes the proof. □

3.1.4 Other norms

Additional to the continuous functions, universal approximation theorems can be shown for various other function classes and topologies, which may also allow for the approximation of functions exhibiting discontinuities or singularities. To give but one example, we next state such a result for Lebesgue spaces on compact sets. The proof is left to the reader, see Exercise 3.26.

Corollary 3.18. *Let $d \in \mathbb{N}$, $L \in \mathbb{N}$, $p \in [1, \infty)$, and let $\sigma \in \mathcal{M}$ not be a polynomial. Then for every $\varepsilon > 0$, every compact $K \subseteq \mathbb{R}^d$, and every $f \in L^p(K)$ there exists $\Phi^{f,\varepsilon} \in \mathcal{N}_d^1(\sigma; L)$ such that*

$$\left(\int_K |f(\boldsymbol{x}) - \Phi(\boldsymbol{x})|^p \, \mathrm{d}\boldsymbol{x} \right)^{1/p} \leq \varepsilon.$$

3.2 Superexpressive activations and Kolmogorov's superposition theorem

In the previous section, we saw that a large class of activation functions allow for universal approximation. However, these results did not provide any insights into the necessary neural network size for achieving a specific accuracy.

Before exploring this topic further in the following chapters, we next present a remarkable result that shows how the required neural network size is significantly influenced by the choice of activation function. The result asserts that, with the appropriate activation function, every $f \in C^0(K)$ on a compact set $K \subseteq \mathbb{R}^d$ can be approximated to *every desired accuracy* $\varepsilon > 0$ using a neural network of size $O(d^2)$; in particular the neural network size is independent of $\varepsilon > 0$, K, and f. We will first discuss the one-dimensional case.

Proposition 3.19. *There exists a continuous activation function $\sigma : \mathbb{R} \to \mathbb{R}$ such that for every compact $K \subseteq \mathbb{R}$, every $\varepsilon > 0$ and every $f \in C^0(K)$ there exists $\Phi(x) = \sigma(wx + b) \in \mathcal{N}_1^1(\sigma; 1, 1)$ such that*

$$\sup_{x \in K} |f(x) - \Phi(x)| < \varepsilon.$$

Proof. Denote by $\tilde{\mathbb{P}}_n$ all polynomials $p(x) = \sum_{j=0}^n q_j x^j$ with rational coefficients, i.e. such that $q_j \in \mathbb{Q}$ for all $j = 0, \dots, n$. Then $\tilde{\mathbb{P}}_n$ can be identified with the n-fold cartesian product $\mathbb{Q} \times \cdots \times \mathbb{Q}$, and thus $\tilde{\mathbb{P}}_n$ is a countable set. Consequently also the set $\tilde{\mathbb{P}} := \bigcup_{n \in \mathbb{N}} \tilde{\mathbb{P}}_n$ of all polynomials with rational coefficients is countable. Let $(p_i)_{i \in \mathbb{Z}}$ be an enumeration of these polynomials, and set

$$\sigma(x) := \begin{cases} p_i(x - 2i) & \text{if } x \in [2i, 2i + 1] \\ p_i(1)(2i + 2 - x) + p_{i+1}(0)(x - 2i - 1) & \text{if } x \in (2i + 1, 2i + 2). \end{cases}$$

In words, σ equals p_i on even intervals $[2i, 2i + 1]$ and is linear on odd intervals $[2i + 1, 2i + 2]$, resulting in a continuous function overall.

We first assume $K = [0, 1]$. By Example 3.5, for every $\varepsilon > 0$ exists $p(x) = \sum_{j=1}^n r_j x^j$ such that $\sup_{x \in [0,1]} |p(x) - f(x)| < \varepsilon/2$. Now choose $q_j \in \mathbb{Q}$ so close to r_j such that $\tilde{p}(x) := \sum_{j=1}^n q_j x^j$ satisfies $\sup_{x \in [0,1]} |\tilde{p}(x) - p(x)| < \varepsilon/2$. Let $i \in \mathbb{Z}$ such that $\tilde{p}(x) = p_i(x)$, i.e., $p_i(x) = \sigma(2i + x)$ for all $x \in [0, 1]$. Then $\sup_{x \in [0,1]} |f(x) - \sigma(x + 2i)| < \varepsilon$.

For general compact K assume that $K \subseteq [a, b]$. By Tietze's extension theorem, f allows a continuous extension to $[a, b]$, so without loss of generality $K = [a, b]$. By the first case we can find $i \in \mathbb{Z}$ such that with $y = (x - a)/(b - a)$ (i.e. $y \in [0, 1]$ if $x \in [a, b]$)

$$\sup_{x \in [a,b]} \left| f(x) - \sigma\left(\frac{x - a}{b - a} + 2i \right) \right| = \sup_{y \in [0,1]} |f(y \cdot (b - a) + a) - \sigma(y - 2i)| < \varepsilon,$$

which gives the statement with $w = 1/(b - a)$ and $b = -a \cdot (b - a) + 2i$. □

To extend this result to arbitrary dimension, we will use Kolmogorov's superposition theorem. It states that every continuous function of d variables can be expressed as a composition of functions that each depend only on one variable. We omit the technical proof, which can be found in [118].

Theorem 3.20 (Kolmogorov)**.** *For every $d \in \mathbb{N}$ exist $2d^2 + d$ monotonically increasing functions $\varphi_{i,j} \in C^0(\mathbb{R})$, $i = 1, \dots, d$, $j = 1, \dots, 2d+1$, such that for every $f \in C^0([0,1]^d)$ there exist functions $f_j \in C^0(\mathbb{R})$, $j = 1, \dots, 2d+1$ satisfying*

$$f(\boldsymbol{x}) = \sum_{j=1}^{2d+1} f_j \left(\sum_{i=1}^{d} \varphi_{i,j}(x_i) \right) \qquad \text{for all } \boldsymbol{x} \in [0,1]^d.$$

Corollary 3.21. *Let $d \in \mathbb{N}$. With the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ from Proposition 3.19, for every compact $K \subseteq \mathbb{R}^d$, every $\varepsilon > 0$ and every $f \in C^0(K)$ there exists $\Phi \in \mathcal{N}_d^1(\sigma; 2, 2d^2 + d)$ (i.e.* $\operatorname{width}(\Phi) = 2d^2 + d$ *and* $\operatorname{depth}(\Phi) = 2$*) such that*

$$\sup_{\boldsymbol{x} \in K} |f(\boldsymbol{x}) - \Phi(\boldsymbol{x})| < \varepsilon.$$

Proof. Without loss of generality we can assume $K = [0,1]^d$: the extension to the general case then follows by Tietze's extension theorem and a scaling argument as in the proof of Proposition 3.19.

Let f_j, $\varphi_{i,j}$, $i = 1, \dots, d$, $j = 1, \dots, 2d+1$ be as in Theorem 3.20. Fix $\varepsilon > 0$. Let $a > 0$ be so large that

$$\sup_{i,j} \sup_{x \in [0,1]} |\varphi_{i,j}(x)| \leq a.$$

Since each f_j is uniformly continuous on the compact set $[-da, da]$, we can find $\delta > 0$ such that

$$\sup_{j} \sup_{\substack{|y-\tilde{y}|<\delta \\ |y|,|\tilde{y}| \leq da}} |f_j(y) - f_j(\tilde{y})| < \frac{\varepsilon}{2(2d+1)}. \tag{3.2.1}$$

By Proposition 3.19 there exist $w_{i,j}$, $b_{i,j} \in \mathbb{R}$ such that

$$\sup_{i,j} \sup_{x \in [0,1]} |\varphi_{i,j}(x) - \underbrace{\sigma(w_{i,j}x + b_{i,j})}_{=:\tilde{\varphi}_{i,j}(x)}| < \frac{\delta}{d} \tag{3.2.2}$$

and w_j, $b_j \in \mathbb{R}$ such that

$$\sup_{j} \sup_{|y| \leq a+\delta} |f_j(y) - \underbrace{\sigma(w_j y + b_j)}_{=:\tilde{f}_j(x)}| < \frac{\varepsilon}{2(2d+1)}. \tag{3.2.3}$$

Then for all $\boldsymbol{x} \in [0,1]^d$ by (3.2.2)

$$\left| \sum_{i=1}^{d} \varphi_{i,j}(x_i) - \sum_{i=1}^{d} \tilde{\varphi}_{i,j}(x_i) \right| < d\frac{\delta}{d} = \delta.$$

Thus with

$$y_j := \sum_{j=1}^{d} \varphi_{i,j}(x_i), \qquad \tilde{y}_j := \sum_{j=1}^{d} \tilde{\varphi}_{i,j}(x_i)$$

it holds $|y_j - \tilde{y}_j| < \delta$. Using (3.2.1) and (3.2.3) we conclude

$$\begin{aligned}
\left| f(\boldsymbol{x}) - \sum_{j=1}^{2d+1} \sigma \left(w_j \cdot \left(\sum_{i=1}^{d} \sigma(w_{i,j} x_i + b_{i,j}) \right) + b_j \right) \right| &= \left| \sum_{j=1}^{2d+1} (f_j(y_j) - \tilde{f}_j(\tilde{y}_j)) \right| \\
&\leq \sum_{j=1}^{2d+1} \Big(|f_j(y_j) - f_j(\tilde{y}_j)| + |f_j(\tilde{y}_j) - \tilde{f}_j(\tilde{y}_j)| \Big) \\
&\leq \sum_{j=1}^{2d+1} \left(\frac{\varepsilon}{2(2d+1)} + \frac{\varepsilon}{2(2d+1)} \right) \leq \varepsilon.
\end{aligned}$$

This concludes the proof. □

Kolmogorov's superposition theorem is intriguing as it shows that approximating d-dimensional functions can be reduced to the (generally much simpler) one-dimensional case through compositions. Neural networks, by nature, are well suited to approximate functions with compositional structures. However, the functions f_j in Theorem 3.20, even though only one-dimensional, could become very complex and challenging to approximate themselves if d is large.

Similarly, the "magic" activation function in Proposition 3.19 encodes the information of all rational polynomials on the unit interval, which is why a neural network of size $O(1)$ suffices to approximate every function to arbitrary accuracy. Naturally, no practical algorithm can efficiently identify appropriate neural network weights and biases for this architecture. As such, the results presented in Section 3.2 should be taken with a pinch of salt as their practical relevance is highly limited. Nevertheless, they highlight that while universal approximation is a fundamental and important property of neural networks, it leaves many aspects unexplored. To gain further insight into practically relevant architectures, in the following chapters, we investigate neural networks with activation functions such as the ReLU.

Bibliography and further reading

The foundation of universal approximation theorems goes back to the late 1980s with seminal works by Cybenko [43], Hornik et al. [94, 93], Funahashi [62] and Carroll and Dickinson [32]. These results were subsequently extended to a wider range of activation functions and architectures. The present analysis in Section 3.1 closely follows the arguments in [130], where it was essentially shown that universal approximation can be achieved if the activation function is not polynomial.

Kolmogorov's superposition theorem stated in Theorem 3.20 was originally proven in 1957 [118]. For a more recent and constructive proof see for instance [26]. Kolmogorov's theorem and its obvious connections to neural networks have inspired various research in this field, e.g. [160, 122, 149, 204, 103], with its practical relevance being debated [67, 121]. The idea for the "magic" activation function in Section 3.2 comes from [138] where it is shown that such an activation function can even be chosen monotonically increasing.

Exercises

Exercise 3.22. Write down a generator of a (minimal) topology on $C^0(\mathbb{R}^d)$ such that $f_n \to f \in C^0(\mathbb{R}^d)$ if and only if $f_n \xrightarrow{\text{cc}} f$, and show this equivalence. This topology is referred to as the topology of compact convergence.

Exercise 3.23. Show the implication "$\Rightarrow$" of Theorem 3.8 and Corollary 3.17.

Exercise 3.24. Prove Lemma 3.12. *Hint*: Consider $\sigma(nx)$ for large $n \in \mathbb{N}$.

Exercise 3.25. Let $k \in \mathbb{N}$, $\sigma \in \mathcal{M}$ and assume that $\sigma * \varphi \in \mathbb{P}_k$ for all $\varphi \in C_c^\infty(\mathbb{R})$. Show that $\sigma \in \mathbb{P}_k$.

Hint: Consider $\psi \in C_c^\infty(\mathbb{R})$ such that $\psi \geq 0$ and $\int_{\mathbb{R}} \psi(x)\,\mathrm{d}x = 1$ and set $\psi_\varepsilon(x) := \psi(x/\varepsilon)/\varepsilon$. Use that away from the discontinuities of σ it holds $\psi_\varepsilon * \sigma(x) \to \sigma(x)$ as $\varepsilon \to 0$. Conclude that σ is piecewise in $\mathbb{P}_k$, and finally show that $\sigma \in C^k(\mathbb{R})$.

Exercise 3.26. Prove Corollary 3.18 with the use of Corollary 3.17.

Exercise 3.27. Complete the proof of Proposition 3.16 for $L > 1$.