Chapter 8

High-dimensional approximation

In the previous chapters we established convergence rates for the approximation of a function $f : [0,1]^d \to \mathbb{R}$ by a neural network. For example, Theorem 7.7 provides the error bound $\mathcal{O}(N^{-(k+s)/d})$ in terms of the network size N (up to logarithmic terms), where k and s describe the smoothness of f. Achieving an accuracy of $\varepsilon > 0$, therefore, necessitates a network size $N = O(\varepsilon^{-d/(k+s)})$ (according to this bound). Hence, the size of the network needs to increase exponentially in d. This exponential dependence on the dimension d is referred to as the **curse of dimensionality** [16]. For classical smoothness spaces, such exponential d dependence cannot be avoided [16, 51, 162]. However, functions f that are of interest in practice may have additional properties, which allow for better convergence rates.

In this chapter, we discuss three scenarios under which the curse of dimensionality can be mitigated. First, we examine an assumption limiting the behavior of functions in their Fourier domain. This assumption allows for slow but dimension independent approximation rates. Second, we consider functions with a specific compositional structure. Concretely, these functions are constructed by compositions and linear combinations of simple low-dimensional subfunctions. In this case, the curse of dimension is present but only through the input dimension of the subfunctions. Finally, we study the situation, where we still approximate high-dimensional functions, but only care about the approximation accuracy on a lower dimensional submanifold. Here, the approximation rate is goverened by the smoothness and the dimension of the manifold.

8.1 The Barron class

In [10], Barron introduced a set of functions that can be approximated by neural networks without a curse of dimensionality. This set, known as the **Barron class**, is characterized by a specific type of bounded variation. To define it, for $f \in L^1(\mathbb{R}^d)$ denote by

$$\hat{f}(\boldsymbol{w}) \coloneqq \int_{\mathbb{R}^d} f(\boldsymbol{x}) e^{-2\pi \mathrm{i}\, \boldsymbol{w}^\top \boldsymbol{x}} \,\mathrm{d}\boldsymbol{x}$$

its Fourier transform. Then, for $C > 0$ the Barron class is defined as

$$\Gamma_C \coloneqq \left\{ f \in L^1(\mathbb{R}^d) \,\middle|\, \|\hat{f}\|_{L^1(\mathbb{R}^d)} < \infty,\ \int_{\mathbb{R}^d} |2\pi\boldsymbol{\xi}||\hat{f}(\boldsymbol{\xi})| \,\mathrm{d}\boldsymbol{\xi} < C \right\}.$$

We point out that the definition of Γ_C in [10] is more general, but our assumption will simplify some of the arguments. Nonetheless, the following proof is very close to the original result, and the presentation is similar to [173, Section 5]. Theorem 1 in [10] reads as follows.

Theorem 8.1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be sigmoidal (see Definition 3.11) and let $f \in \Gamma_C$ for some $C > 0$. Denote by $B_1^d := \{\boldsymbol{x} \in \mathbb{R}^d \,|\, \|\boldsymbol{x}\| \leq 1\}$ the unit ball. Then, for every $c > 4C^2$ and every $N \in \mathbb{N}$ there exists a neural network Φ^f with architecture $(\sigma; d, N, 1)$ such that*

$$\frac{1}{|B_1^d|} \int_{B_1^d} \left| f(\boldsymbol{x}) - \Phi^f(\boldsymbol{x}) \right|^2 \mathrm{d}\boldsymbol{x} \leq \frac{c}{N}, \tag{8.1.1}$$

where $|B_1^d|$ is the Lebesgue measure of B_1^d.

Remark 8.2. The approximation rate on (8.1.1) can be slightly improved under some assumptions on the activation function such as powers of the ReLU, [212].

Importantly, the dimension d does not enter on the right-hand side of (8.1.1), in particular the convergence rate is not directly affected by the dimension, which is in stark contrast to the results of the previous chapters. However, it should be noted, that the constant C_f may still have some inherent d-dependence, see Exercise 8.10.

The proof of Theorem 8.1 is based on a peculiar property of high-dimensional convex sets, which is described by the (approximate) Caratheodory theorem, the original version of which was given in [30]. The more general version stated in the following lemma follows [235, Theorem 0.0.2] and [10, 175]. For its statement recall that $\overline{\mathrm{co}}(G)$ denotes the the closure of the convex hull of G.

Lemma 8.3. *Let H be a Hilbert space, and let $G \subseteq H$ be such that for some $B > 0$ it holds that $\|g\|_H \leq B$ for all $g \in G$. Let $f \in \overline{\mathrm{co}}(G)$. Then, for every $N \in \mathbb{N}$ and every $c > B^2$ there exist $(g_i)_{i=1}^N \subseteq G$ such that*

$$\left\| f - \frac{1}{N} \sum_{i=1}^N g_i \right\|_H^2 \leq \frac{c}{N}. \tag{8.1.2}$$

Proof. Fix $\varepsilon > 0$ and $N \in \mathbb{N}$. Since $f \in \overline{\mathrm{co}}(G)$, there exist coefficients $\alpha_1, \dots, \alpha_m \in [0, 1]$ summing to 1, and linearly independent elements $h_1, \dots, h_m \in G$ such that

$$f^* := \sum_{j=1}^m \alpha_j h_j$$

satisfies $\|f - f^*\|_H < \varepsilon$. We claim that there exists $g_1, \dots, g_N$, each in $\{h_1, \dots, h_m\}$, such that

$$\left\| f^* - \frac{1}{N} \sum_{j=1}^N g_j \right\|_H^2 \leq \frac{B^2}{N}. \tag{8.1.3}$$

Since $\varepsilon > 0$ was arbitrary, this then concludes the proof. Since there exists an isometric isomorphism from $\operatorname{span}\{h_1, \dots, h_m\}$ to $\mathbb{R}^m$, there is no loss of generality in assuming $H = \mathbb{R}^m$ in the following.

Let X_i, $i = 1, \dots, N$, be i.i.d. $\mathbb{R}^m$-valued random variables with

$$\mathbb{P}[X_i = h_j] = \alpha_j \qquad \text{for all } i = 1, \dots, m.$$

In particular $\mathbb{E}[X_i] = \sum_{j=1}^m \alpha_j h_j = f^*$ for each i. Moreover,

$$\begin{aligned}
\mathbb{E}\left[\left\| f^* - \frac{1}{N}\sum_{j=1}^N X_j \right\|^2\right] &= \mathbb{E}\left[\left\| \frac{1}{N}\sum_{j=1}^N (f^* - X_j) \right\|^2\right] \\
&= \frac{1}{N^2}\left[\sum_{j=1}^N \|f^* - X_j\|^2 + \sum_{i\neq j} \langle f^* - X_i, f^* - X_j\rangle\right] \\
&= \frac{1}{N}\mathbb{E}[\|f^* - X_1\|^2] \\
&= \frac{1}{N}\mathbb{E}[\|f^*\| - 2\langle f^*, X_1\rangle + \|X_1\|^2] \\
&= \frac{1}{N}\mathbb{E}[\|X_1\|^2 - \|f^*\|^2] \leq \frac{B^2}{N}. \qquad (8.1.4)
\end{aligned}$$

Here we used that the $(X_i)_{i=1}^N$ are independent, the fact that $\mathbb{E}[X_i] = f^*$, as well as $\mathbb{E}\langle f^* - X_i, f^* - X_j\rangle = 0$ if $i \neq j$. Since the expectation in (8.1.4) is bounded by B^2/N, there must exist at least one realization of the random variables $X_i \in \{h_1, \dots, h_m\}$, denoted as g_i, for which (8.1.3) holds. □

Lemma 8.3 provides a powerful tool: If we want to approximate a function f with a superposition of N elements in a set G, then it is sufficient to show that f can be represented as an arbitrary (infinite) convex combination of elements of G.

Lemma 8.3 suggests that we can prove Theorem 8.1 by showing that each function in Γ_C belongs to the convex hull of neural networks with just a single neuron. We make a small detour before proving this result. We first show that each function $f \in \Gamma_C$ is in the convex hull of affine transforms of Heaviside functions. We define the *set of affine transforms of Heaviside functions* G_C as

$$G_C \coloneqq \left\{ B_1^d \ni \boldsymbol{x} \mapsto \gamma \cdot \mathbb{1}_{\mathbb{R}_+}(\langle \boldsymbol{a}, \boldsymbol{x}\rangle + b) \,\middle|\, \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}, |\gamma| \leq 2C \right\}.$$

The following lemma, corresponding to [173, Lemma 5.12], provides a link between Γ_C and G_C.

Lemma 8.4. *Let $d \in \mathbb{N}$, $C > 0$ and $f \in \Gamma_C$. Then $f|_{B_1^d} - f(0) \in \overline{\operatorname{co}}(G_C)$, where the closure is taken with respect to the norm*

$$\|g\|_{L^{2,\diamond}(B_1^d)} \coloneqq \left(\frac{1}{|B_1^d|}\int_{B_1^d} |g(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x}\right)^{1/2}.$$

Proof. Since $f \in \Gamma_C$, we have that $f,\ \hat{f} \in L^1(\mathbb{R}^d)$. Hence, we can apply the inverse Fourier transform and get the following computation:

$$\begin{aligned} f(\boldsymbol{x}) - f(0) &= \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi}) \left(e^{2\pi i \langle \boldsymbol{x}, \boldsymbol{\xi} \rangle} - 1 \right) \mathrm{d}\boldsymbol{\xi} \\ &= \int_{\mathbb{R}^d} \left| \hat{f}(\boldsymbol{\xi}) \right| \left(e^{2\pi i \langle \boldsymbol{x}, \boldsymbol{\xi} \rangle + i\kappa(\boldsymbol{\xi})} - e^{i\kappa(\boldsymbol{\xi})} \right) \mathrm{d}\boldsymbol{\xi} \\ &= \int_{\mathbb{R}^d} \left| \hat{f}(\boldsymbol{\xi}) \right| \left(\cos(2\pi \langle \boldsymbol{x}, \boldsymbol{\xi} \rangle + \kappa(\boldsymbol{\xi})) - \cos(\kappa(\boldsymbol{\xi})) \right) \mathrm{d}\boldsymbol{\xi}, \end{aligned}$$

where $\kappa(\boldsymbol{\xi})$ is the phase of $\hat{f}(\boldsymbol{\xi})$ and the last inequality follows since f is real-valued.

To use the fact that f has a bounded Fourier moment, we reformulate the integral as

$$\begin{aligned} &\int_{\mathbb{R}^d} \left| \hat{f}(\boldsymbol{\xi}) \right| \left(\cos(2\pi \langle x, \boldsymbol{\xi} \rangle + \kappa(\boldsymbol{\xi})) - \cos(\kappa(\boldsymbol{\xi})) \right) \mathrm{d}\boldsymbol{\xi} \\ &\quad = \int_{\mathbb{R}^d} \frac{\left(\cos(2\pi \langle x, \boldsymbol{\xi} \rangle + \kappa(\boldsymbol{\xi})) - \cos(\kappa(\boldsymbol{\xi})) \right)}{|2\pi\boldsymbol{\xi}|} |2\pi\boldsymbol{\xi}| \left| \hat{f}(\boldsymbol{\xi}) \right| \mathrm{d}\boldsymbol{\xi}. \end{aligned}$$

We define a new measure Λ with density

$$\mathrm{d}\Lambda(\boldsymbol{\xi}) \coloneqq \frac{1}{C} |2\pi\boldsymbol{\xi}| |\hat{f}(\boldsymbol{\xi})| \, \mathrm{d}\boldsymbol{\xi}.$$

Since $f \in \Gamma_C$, it follows that Λ is a probability measure on $\mathbb{R}^d$. Now we have that

$$f(\boldsymbol{x}) - f(0) = C \int_{\mathbb{R}^d} \frac{\left(\cos(2\pi \langle \boldsymbol{x}, \boldsymbol{\xi} \rangle + \kappa(\boldsymbol{\xi})) - \cos(\kappa(\boldsymbol{\xi})) \right)}{|2\pi\boldsymbol{\xi}|} \, \mathrm{d}\Lambda(\boldsymbol{\xi}). \tag{8.1.5}$$

Next, we would like to replace the integral of (8.1.5) by an appropriate finite sum.

The cosine function is 1-Lipschitz. Hence, we note that $\boldsymbol{\xi} \mapsto q_{\boldsymbol{x}}(\boldsymbol{\xi}) \coloneqq (\cos(2\pi \langle \boldsymbol{x}, \boldsymbol{\xi} \rangle + \kappa(\boldsymbol{\xi})) - \cos(\kappa(\boldsymbol{\xi})))/|2\pi\boldsymbol{\xi}|$ is bounded by 1. In addition, it is easy to see that $q_{\boldsymbol{x}}$ is well-defined and continuous even in the origin.

Therefore, the integral (8.1.5) can be approximated by a Riemann sum, i.e.,

$$\left| C \int_{\mathbb{R}^d} q_{\boldsymbol{x}}(\boldsymbol{\xi}) \, \mathrm{d}\Lambda(\boldsymbol{\xi}) - C \sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} q_{\boldsymbol{x}}(\theta) \cdot \Lambda(I_\theta) \right| \to 0, \tag{8.1.6}$$

where $I_\theta \coloneqq [0, 1/n)^d + \theta$.

Since $f(\boldsymbol{x}) - f(0)$ is continuous and thus bounded on B_1^d, we have by the dominated convergence theorem that

$$\frac{1}{|B_1^d|} \int_{B_1^d} \left| f(\boldsymbol{x}) - f(0) - C \sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} q_{\boldsymbol{x}}(\theta) \cdot \Lambda(I_\theta) \right|^2 \mathrm{d}\boldsymbol{x} \to 0. \tag{8.1.7}$$

Since $\sum_{\theta \in \frac{1}{n}\mathbb{Z}^d} \Lambda(I_\theta) = \Lambda(\mathbb{R}^d) = 1$, we conclude that $f(\boldsymbol{x}) - f(0)$ is in the $L^{2,\diamond}(B_1^d)$ closure of convex combinations of functions of the form

$$\boldsymbol{x} \mapsto g_\theta(\boldsymbol{x}) \coloneqq \alpha_\theta q_{\boldsymbol{x}}(\theta),$$

for $\theta \in \mathbb{R}^d$ and $0 \leq \alpha_\theta \leq C$.

Now we only need to prove that each g_θ is in $\overline{\mathrm{co}}(G_C)$. By setting $z = \langle \boldsymbol{x}, \theta/|\theta| \rangle$, we observe that the result follows if the map

$$[-1,1] \ni z \mapsto \alpha_\theta \frac{\cos(2\pi|\theta|z + \kappa(\theta)) - \cos(\kappa(\theta))}{|2\pi\theta|} =: \tilde{g}_\theta(z),$$

can be approximated arbitrarily well by convex combinations of functions of the form

$$[-1,1] \ni z \mapsto \gamma \mathbb{1}_{\mathbb{R}_+}\left(a'z + b'\right), \tag{8.1.8}$$

where a', $b' \in \mathbb{R}$ and $|\gamma| \leq 2C$.

We define, for $T \in \mathbb{N}$,

$$g_{T,+} \coloneqq \sum_{i=1}^{T} \frac{\left|\tilde{g}_\theta\left(\frac{i}{T}\right) - \tilde{g}_\theta\left(\frac{i-1}{T}\right)\right|}{2C} \left(2C\mathrm{sign}\left(\tilde{g}_\theta\left(\frac{i}{T}\right) - \tilde{g}_\theta\left(\frac{i-1}{T}\right)\right) \mathbb{1}_{\mathbb{R}_+}\left(x - \frac{i}{T}\right)\right),$$

$$g_{T,-} \coloneqq \sum_{i=1}^{T} \frac{\left|\tilde{g}_\theta\left(-\frac{i}{T}\right) - \tilde{g}_\theta\left(\frac{1-i}{T}\right)\right|}{2C} \left(2C\mathrm{sign}\left(\tilde{g}_\theta\left(-\frac{i}{T}\right) - \tilde{g}_\theta\left(\frac{1-i}{T}\right)\right) \mathbb{1}_{\mathbb{R}_+}\left(-x + \frac{i}{T}\right)\right).$$

Per construction, $g_{T,-} + g_{T,+}$ converges to $\tilde{g}_\theta$ for $T \to \infty$. Moreover, $\|\tilde{g}_\theta'\|_{L^\infty(\mathbb{R})} \leq C$ and hence

$$\begin{aligned}
&\sum_{i=1}^{T} \frac{|\tilde{g}_\theta(i/T) - \tilde{g}_\theta((i-1)/T)|}{2C} + \sum_{i=1}^{T} \frac{|\tilde{g}_\theta(-i/T) - \tilde{g}_\theta((1-i)/T)|}{2C} \\
&\leq \frac{2}{2CT} \sum_{i=1}^{T} \|\tilde{g}_\theta'\|_{L^\infty(\mathbb{R})} \leq 1.
\end{aligned}$$

We conclude that $g_{T,-} + g_{T,+}$ is a convex combination of functions of the form (8.1.8). Hence, $\tilde{g}_\theta$ can be arbitrarily well approximated by convex combinations of the form (8.1.8). Therefore $g_\theta \in \overline{\mathrm{co}}(G_C)$. Finally, (8.1.7) yields that $f - f(0) \in \overline{\mathrm{co}}(G_C)$. □

We now have all tools to complete the proof of Theorem 8.1.

of Theorem 8.1. Let $f \in \Gamma_C$. By Lemma 8.4

$$f|_{B_1^d} - f(0) \in \overline{\mathrm{co}}(G_C).$$

It is not hard to see that for every $g \in G_C$ holds $\|g\|_{L^{2,\diamond}(B_1^d)} \leq 2C$. Applying Lemma 8.3 with the Hilbert space $L^{2,\diamond}(B_1^d)$, we get that for every $N \in \mathbb{N}$ there exist $|\gamma_i| \leq 2C$, $\boldsymbol{a}_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$, for $i = 1, \ldots, N$, so that

$$\frac{1}{|B_1^d|} \int_{B_1^d} \left| f(\boldsymbol{x}) - f(0) - \sum_{i=1}^{N} \gamma_i \mathbb{1}_{\mathbb{R}_+}(\langle \boldsymbol{a}_i, x \rangle + b_i) \right|^2 \mathrm{d}\boldsymbol{x} \leq \frac{4C^2}{N}.$$

By Exercise 3.24, it holds that $\sigma(\lambda \cdot) \to \mathbb{1}_{\mathbb{R}_+}$ for $\lambda \to \infty$ almost everywhere. Thus, for every $\delta > 0$ there exist $\tilde{\boldsymbol{a}}_i$, $\tilde{b}_i$, $i = 1, \ldots, N$, so that

$$\frac{1}{|B_1^d|}\int_{B_1^d}\left|f(\boldsymbol{x})-f(0)-\sum_{i=1}^{N}\gamma_i\sigma\left(\langle\tilde{\boldsymbol{a}}_i,\boldsymbol{x}\rangle+\tilde{b}_i\right)\right|^2\,\mathrm{d}\boldsymbol{x}\leq\frac{4C^2}{N}+\delta.$$

The result follows by observing that

$$\sum_{i=1}^{N}\gamma_i\sigma\left(\langle\tilde{\boldsymbol{a}}_i,\boldsymbol{x}\rangle+\tilde{b}_i\right)+f(0)$$

is a neural network with architecture $(\sigma;d,N,1)$. □

The dimension-independent approximation rate of Theorem 8.1 may seem surprising, especially in comparison to the results in Chapters 4 and 5. However, this can be explained by recognizing that the assumption of a finite Fourier moment is effectively a *dimension-dependent regularity assumption*. Indeed, the condition becomes more restrictive in higher dimensions and hence the complexity of Γ_C does not grow with the dimension.

To further explain this, let us relate the Barron class to classical function spaces. In [10, Section II] it was observed that a sufficient condition is that all derivatives of order up to $\lfloor d/2\rfloor+2$ are square-integrable. In other words, if f belongs to the Sobolev space $H^{\lfloor d/2\rfloor+2}(\mathbb{R}^d)$, then f is a Barron function. Importantly, the functions must become smoother, as the dimension increases. This assumption would also imply an approximation rate of $N^{-1/2}$ in the L^2 norm by sums of at most N B-splines, see [166, 51]. However, in such estimates some constants may still depend exponentially on d, whereas all constants in Theorem 8.1 are controlled independently of d.

Another notable aspect of the approximation of Barron functions is that the absolut values of the weights other than the output weights are not bounded by a constant. To see this, we refer to (8.1.6), where arbitrarily large θ need to be used. While Γ_C is a compact set, the set of neural networks of the specified architecture for a fixed $N\in\mathbb{N}$ is not parameterized with a compact parameter set. In a certain sense, this is reminiscent of Proposition 3.19 and Theorem 3.20, where arbitrarily strong approximation rates where achieved by using a very complex activation function and a non-compact parameter space.

8.2 Functions with compositionality structure

As a next instance of types of functions for which the curse of dimensionality can be overcome, we study functions with compositional structure. In words, this means that we study high-dimensional functions that are constructed by composing many low-dimensional functions. This point of view was proposed in [176]. Note that this can be a realistic assumption in many cases, such as for sensor networks, where local information is first aggregated in smaller clusters of sensors before some information is sent to a processing unit for further evaluation.

We introduce a model for compositional functions next. Consider a directed acyclic graph $\mathcal{G}$ with M vertices $\eta_1,\dots,\eta_M$ such that

- exactly d vertices, $\eta_1,\dots,\eta_d$, have no ingoing edge,
- each vertex has at most $m\in\mathbb{N}$ ingoing edges,
- exactly one vertex, η_M, has no outgoing edge.

With each vertex η_j for $j > d$ we associate a function $f_j : \mathbb{R}^{d_j} \to \mathbb{R}$. Here d_j denotes the cardinality of the set S_j, which is defined as the set of indices i corresponding to vertices η_i for which we have an edge from η_i to η_j. Without loss of generality, we assume that $m \geq d_j = |S_j| \geq 1$ for all $j > d$. Finally, we let

$$F_j := x_j \quad \text{for all} \quad j \leq d \tag{8.2.1a}$$

and[1]

$$F_j := f_j((F_i)_{i \in S_j}) \quad \text{for all} \quad j > d. \tag{8.2.1b}$$

Then $F_M(x_1, \dots, x_d)$ is a function from $\mathbb{R}^d \to \mathbb{R}$. Assuming

$$\|f_j\|_{C^{k,s}(\mathbb{R}^{d_j})} \leq 1 \quad \text{for all} \quad j = d+1, \dots, M, \tag{8.2.2}$$

we denote the set of all functions of the type F_M by $\mathcal{F}^{k,s}(m, d, M)$. Figure 8.1 shows possible graphs of such functions.

Clearly, for $s = 0$, $\mathcal{F}^{k,0}(m, d, M) \subseteq C^k(\mathbb{R}^d)$ since the composition of functions in C^k belongs again to C^k. A direct application of Theorem 7.7 allows to approximate $F_M \in \mathcal{F}^k(m, d, M)$ with a neural network of size $O(N \log(N))$ and error $O(N^{-\frac{k}{d}})$. Since each f_j depends only on m variables, intuitively we expect an error convergence of type $O(N^{-\frac{k}{m}})$ with the constant somehow depending on the number M of vertices. To show that this is actually possible, in the following we associate with each node η_j a depth $l_j \geq 0$, such that l_j is the maximum number of edges connecting η_j to one of the nodes $\{\eta_1, \dots, \eta_d\}$.

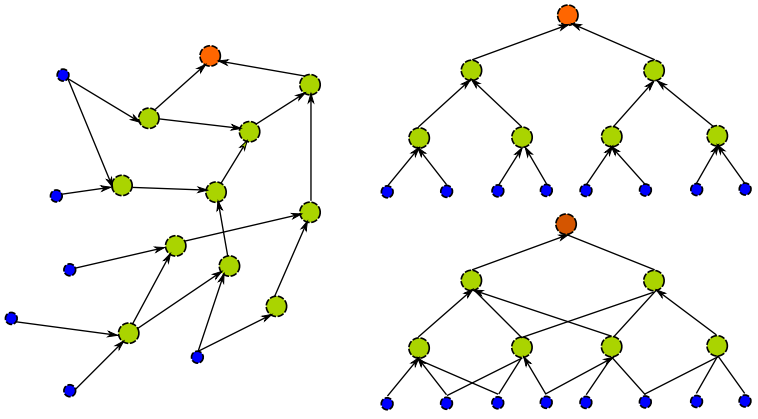

Figure 8.1: Three types of graphs that could be the basis of compositional functions. The associated functions are composed of two or three-dimensional functions only.

Proposition 8.5. *Let k, m, d, $M \in \mathbb{N}$ and $s > 0$. Let $F_M \in \mathcal{F}^{k,s}(m, d, M)$. Then there exists a constant $C = C(m, k+s, M)$ such that for every $N \in \mathbb{N}$ there exists a ReLU neural network $\hat{F}_M$*

[1]The ordering of the inputs $(F_i)_{i \in S_j}$ in (8.2.1b) is arbitrary but considered fixed throughout.

such that

$$\operatorname{size}(\hat{F}_M) \leq CN\log(N), \qquad \operatorname{depth}(\hat{F}_M) \leq C\log(N)$$

and

$$\sup_{\boldsymbol{x}\in[0,1]^d} |F_M(\boldsymbol{x}) - \hat{F}_M(\boldsymbol{x})| \leq N^{-\frac{k+s}{m}}.$$

Proof. Throughout this proof we assume without loss of generality that the indices follow a topological ordering, i.e., they are ordered such that $S_j \subseteq \{1,\dots,j-1\}$ for all j (i.e. the inputs of vertex η_j can only be vertices η_i with $i<j$).

Step 1. First assume that there exists functions $\hat{f}_j$ such that

$$|f_j(\boldsymbol{x}) - \hat{f}_j(\boldsymbol{x})| \leq \delta_j := \varepsilon\cdot(2m)^{-(M+1-j)} \quad \text{for all} \quad \boldsymbol{x}\in[-2,2]^{d_j}. \tag{8.2.3}$$

Let $\hat{F}_j$ be defined as in (8.2.1), but with all f_j in (8.2.1b) replaced by $\hat{f}_j$. We now check the error of the approximation $\hat{F}_M$ to F_M. To do so we proceed by induction over j and show that for all $\boldsymbol{x}\in[-1,1]^d$

$$|F_j(\boldsymbol{x}) - \hat{F}_j(\boldsymbol{x})| \leq (2m)^{-(M-j)}\varepsilon. \tag{8.2.4}$$

Note that due to $\|f_j\|_{C^k} \leq 1$ we have $|F_j(\boldsymbol{x})| \leq 1$ and thus (8.2.4) implies in particular that $\hat{F}_j(\boldsymbol{x}) \in [-2,2]$.

For $j=1$ it holds $F_1(x_1) = \hat{F}_1(x_1) = x_1$, and thus (8.2.4) is valid for all $x_1\in[-1,1]$. For the induction step, for all $\boldsymbol{x}\in[-1,1]^d$ by (8.2.3) and the induction hypothesis

$$\begin{aligned}
|F_j(\boldsymbol{x}) - \hat{F}_j(\boldsymbol{x})| &= |f_j((F_i)_{i\in S_j}) - \hat{f}_j((\hat{F}_i)_{i\in S_j})| \\
&= |f_j((F_i)_{i\in S_j}) - f_j((\hat{F}_i)_{i\in S_j})| + |f_j((\hat{F}_i)_{i\in S_j}) - \hat{f}_j((\hat{F}_i)_{i\in S_j})| \\
&\leq \sum_{i\in S_j} |F_i - \hat{F}_i| + \delta_j \\
&\leq m\cdot(2m)^{-(M-(j-1))}\varepsilon + (2m)^{-(M+1-j)}\varepsilon \\
&\leq (2m)^{-(M-j)}\varepsilon.
\end{aligned}$$

Here we used that $|\frac{d}{dx_r} f_j((x_i)_{i\in S_j})| \leq 1$ for all $r\in S_j$ so that

$$\begin{aligned}
|f_j((x_i)_{i\in S_j}) - f_j((y_i)_{i\in S_j})| &\leq \sum_{r\in S_j} |f((x_i)_{\substack{i\in S_j\\ i\leq r}}, (y_i)_{\substack{i\in S_j\\ i>r}}) - f((x_i)_{\substack{i\in S_j\\ i<r}}, (y_i)_{\substack{i\in S_j\\ i\geq r}})| \\
&\leq \sum_{r\in S_j} |x_r - y_r|.
\end{aligned}$$

This shows that (8.2.4) holds, and thus for all $\boldsymbol{x}\in[-1,1]^d$

$$|F_M(\boldsymbol{x}) - \hat{F}_M(\boldsymbol{x})| \leq \varepsilon.$$

Step 2. We sketch a construction, of how to write $\hat{F}_M$ from Step 1 as a neural network of the claimed size and depth bounds. Fix $N \in \mathbb{N}$ and let

$$N_j := \lceil N(2m)^{\frac{m}{k+s}(M+1-j)} \rceil.$$

By Theorem 7.7, since $d_j \leq m$, we can find a neural network $\hat{f}_j$ satisfying

$$\sup_{\boldsymbol{x} \in [-2,2]^{d_j}} |f_j(\boldsymbol{x}) - \hat{f}_j(\boldsymbol{x})| \leq N_j^{-\frac{k+s}{m}} \leq N^{-\frac{k+s}{m}} (2m)^{-(M+1-j)} \tag{8.2.5}$$

and

$$\mathrm{size}(\hat{f}_j) \leq C N_j \log(N_j) \leq C N (2m)^{\frac{m(M+1-j)}{k+s}} \left(\log(N) + \log(2m) \frac{m(M+1-j)}{k+s} \right)$$

as well as

$$\mathrm{depth}(\hat{f}_j) \leq C \cdot \left(\log(N) + \log(2m) \frac{m(M+1-j)}{k+s} \right).$$

Then

$$\begin{aligned} \sum_{j=1}^{n} \mathrm{size}(\hat{f}_j) \leq 2CN \log(N) \sum_{j=1}^{M} (2m)^{\frac{m(M+1-j)}{k+s}} &\leq 2CN \log(N) \sum_{j=1}^{M} \left((2m)^{\frac{m}{k+s}} \right)^j \\ &\leq 2CN \log(N) (2m)^{\frac{m(M+1)}{k+s}}. \end{aligned}$$

Here we used $\sum_{j=1}^{M} a^j \leq \int_1^{M+1} \exp(\log(a)x) \, \mathrm{d}x \leq \frac{1}{\log(a)} a^{M+1}$.

The function $\hat{F}_M$ from Step 1 then will yield error $N^{-\frac{k+s}{m}}$ by (8.2.3) and (8.2.5). We observe that $\hat{F}_M$ can be constructed as a neural network by propagating all values $\hat{F}_1, \ldots, \hat{F}_j$ to all consecutive layers using identity neural networks and then using the values $(\hat{F}_i)_{i \in S_{j+1}}$ as input to $\hat{f}_{j+1}$. The depth of this neural network is bounded by

$$\sum_{j=1}^{M} \mathrm{depth}(\hat{f}_j) = O(M \log(N)).$$

We have at most $\sum_{j=1}^{M} |S_j| \leq mM$ values which need to be propagated through these $O(M \log(N))$ layers, amounting to an overhead $O(mM^2 \log(N)) = O(\log(N))$ for the identity neural networks. In all the neural network size is thus $O(N \log(N))$. □

Remark 8.6. From the proof we observe that the constant C in Proposition 8.5 behaves like $O((2m)^{\frac{m(M+1)}{k+s}})$.

8.3 Functions on manifolds

Another instance in which the curse of dimension can be mitigated, is if the input to the network belongs to $\mathbb{R}^d$, but stems from an m-dimensional manifold $\mathcal{M} \subseteq \mathbb{R}^d$. If we only measure the

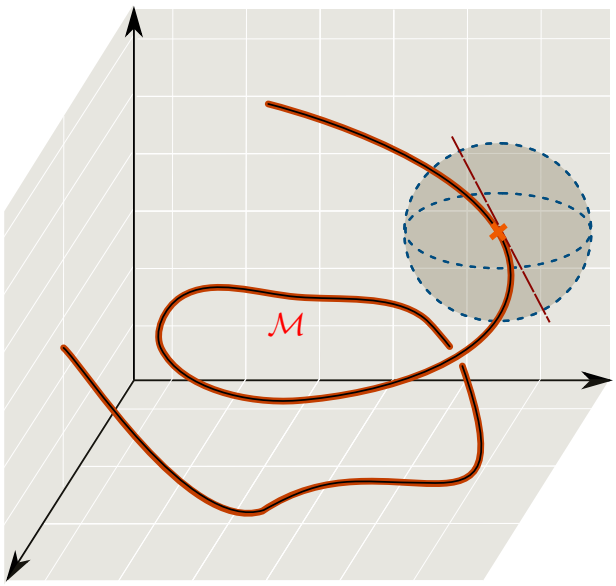

Figure 8.2: One-dimensional sub-manifold of three-dimensional space. At the orange point, we depict a ball and the tangent space of the manifold.

approximation error on $\mathcal{M}$, then we can again show that it is m rather than d that determines the rate of convergence.

To explain the idea, we assume in the following that $\mathcal{M}$ is a smooth, compact m-dimensional manifold in $\mathbb{R}^d$. Moreover, we suppose that there exists $\delta > 0$ and finitely many points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M \in \mathcal{M}$ such that the δ-balls $B_{\delta/2}(\boldsymbol{x}_i) := \{\boldsymbol{y} \in \mathbb{R}^d \,|\, \|\boldsymbol{y} - \boldsymbol{x}\|_2 < \delta/2\}$ for $j = 1, \dots, M$ cover $\mathcal{M}$ (for every $\delta > 0$ such $\boldsymbol{x}_i$ exist since $\mathcal{M}$ is compact). Moreover, denoting by $T_{\boldsymbol{x}}\mathcal{M} \simeq \mathbb{R}^m$ the tangential space of $\mathcal{M}$ at $\boldsymbol{x}$, we assume $\delta > 0$ to be so small that the orthogonal projection

$$\pi_j : B_\delta(\boldsymbol{x}_j) \cap \mathcal{M} \to T_{\boldsymbol{x}_j}\mathcal{M} \tag{8.3.1}$$

is injective, the set $\pi_j(B_\delta(\boldsymbol{x}_j) \cap \mathcal{M}) \subseteq T_{\boldsymbol{x}_j}\mathcal{M}$ has C^∞ boundary, and the inverse projection

$$\pi_j^{-1} : \pi_j(B_\delta(\boldsymbol{x}_j) \cap \mathcal{M}) \to \mathcal{M} \tag{8.3.2}$$

is C^∞ (this is possible because $\mathcal{M}$ is a smooth manifold). A visualization of this assumption is shown in Figure 8.2.

Note that π_j in (8.3.1) is a linear map, whereas π_j^{-1} in (8.3.2) is in general non-linear.

For a function $f : \mathcal{M} \to \mathbb{R}$ and $\boldsymbol{x} \in B_\delta(\boldsymbol{x}_j) \cap \mathcal{M}$ we can then write

$$f(\boldsymbol{x}) = f(\pi_j^{-1}(\pi_j(\boldsymbol{x}))) = f_j(\pi_j(\boldsymbol{x}))$$

where

$$f_j := f \circ \pi_j^{-1} : \pi_j(B_\delta(\boldsymbol{x}_j) \cap \mathcal{M}) \to \mathbb{R}.$$

In the following, for $f : \mathcal{M} \to \mathbb{R}$, $k \in \mathbb{N}_0$, and $s \in [0, 1)$ we let

$$\|f\|_{C^{k,s}(\mathcal{M})} := \sup_{j=1,\dots,M} \|f_j\|_{C^{k,s}(\pi_j(B_\delta(\boldsymbol{x}_j)\cap\mathcal{M}))}.$$

We now state the main result of this section.

Proposition 8.7. *Let $d, k \in \mathbb{N}$, $s \geq 0$, and let $\mathcal{M}$ be a smooth, compact m-dimensional manifold in $\mathbb{R}^d$. Then there exists a constant $C > 0$ such that for all $f \in C^{k,s}(\mathcal{M})$ and every $N \in \mathbb{N}$ there exists a ReLU neural network Φ_N^f such that* $\operatorname{size}(\Phi_N^f) \leq CN\log(N)$, $\operatorname{depth}(\Phi_N^f) \leq C\log(N)$ *and*

$$\sup_{\boldsymbol{x}\in\mathcal{M}} |f(\boldsymbol{x}) - \Phi_N^f(\boldsymbol{x})| \leq C\|f\|_{C^{k,s}(\mathcal{M})} N^{-\frac{k+s}{m}}.$$

Proof. Since $\mathcal{M}$ is compact there exists $A > 0$ such that $\mathcal{M} \subseteq [-A, A]^d$. Similar as in the proof of Theorem 7.7, we consider a uniform mesh with nodes $\{-A + 2A\frac{\boldsymbol{\nu}}{n} \,|\, \boldsymbol{\nu} \leq n\}$, and the corresponding piecewise linear basis functions forming the partition of unity $\sum_{\boldsymbol{\nu}\leq} \varphi_{\boldsymbol{\nu}} \equiv 1$ on $[-A, A]^d$ where $\operatorname{supp}\varphi_{\boldsymbol{\nu}} \leq \{\boldsymbol{y} \in \mathbb{R}^d \,|\, \|\frac{\boldsymbol{\nu}}{n} - \boldsymbol{y}\|_\infty \leq \frac{A}{n}\}$. Let $\delta > 0$ be such as in the beginning of this section. Since $\mathcal{M}$ is covered by the balls $(B_{\delta/2}(\boldsymbol{x}_j))_{j=1}^M$, fixing $n \in \mathbb{N}$ large enough, for each $\boldsymbol{\nu}$ such that $\operatorname{supp}\varphi_{\boldsymbol{\nu}} \cap \mathcal{M} \neq \emptyset$ there exists $j(\boldsymbol{\nu}) \in \{1, \dots, M\}$ such that $\operatorname{supp}\varphi_{\boldsymbol{\nu}} \subseteq B_\delta(\boldsymbol{x}_{j(\boldsymbol{\nu})})$ and we set $I_j := \{\boldsymbol{\nu} \leq M \,|\, j = j(\boldsymbol{\nu})\}$. Then we have for all $\boldsymbol{x} \in \mathcal{M}$

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\nu}\leq n} \varphi_{\boldsymbol{\nu}}(\boldsymbol{x}) f_j(\pi_j(\boldsymbol{x})) = \sum_{j=1}^M \sum_{\boldsymbol{\nu}\in I_j} \varphi_{\boldsymbol{\nu}}(\boldsymbol{x}) f_j(\pi_j(\boldsymbol{x})). \tag{8.3.3}$$

Next, we approximate the functions f_j. Let C_j be the smallest (m-dimensional) cube in $T_{\boldsymbol{x}_j}\mathcal{M} \simeq \mathbb{R}^m$ such that $\pi_j(B_\delta(\boldsymbol{x}_j) \cap \mathcal{M}) \subseteq C_j$. The function $\hat{f}_j$ can be extended to a function on C_j (we will use the same notation for this extension) such that

$$\|f\|_{C^{k,s}(C_j)} \leq C\|f\|_{C^{k,s}(\pi_j(B_\delta(\boldsymbol{x}_j)\cap\mathcal{M}))},$$

for some constant depending on $\pi_j(B_\delta(\boldsymbol{x}_j) \cap \mathcal{M})$ but independent of f. Such an extension result can, for example, be found in [215, Chapter VI]. By Theorem 7.7 (also see Remark 7.9), there exists a neural network $\hat{f}_j : C_j \to \mathbb{R}$ such that

$$\sup_{\boldsymbol{x}\in C_j} |f_j(\boldsymbol{x}) - \hat{f}_j(\boldsymbol{x})| \leq CN^{-\frac{k+s}{m}} \tag{8.3.4}$$

and

$$\operatorname{size}(\hat{f}_j) \leq CN\log(N), \qquad \operatorname{depth}(\hat{f}_j) \leq C\log(N).$$

To approximate f in (8.3.3) we now let with $\varepsilon := N^{-\frac{k+s}{d}}$

$$\Phi_N := \sum_{j=1}^M \sum_{\boldsymbol{\nu}\in I_j} \Phi_\varepsilon^\times(\varphi_{\boldsymbol{\nu}}, \hat{f}_i \circ \pi_j),$$

where we note that π_j is linear and thus $\hat{f}_j \circ \pi_j$ can be expressed by a neural network. First let us estimate the error of this approximation. For $\boldsymbol{x} \in \mathcal{M}$

$$\begin{aligned}
|f(\boldsymbol{x}) - \Phi_N(\boldsymbol{x})| &\leq \sum_{j=1}^{M} \sum_{\boldsymbol{\nu} \in I_j} |\varphi_{\boldsymbol{\nu}}(\boldsymbol{x}) f_j(\pi_j(\boldsymbol{x})) - \Phi_\varepsilon^\times(\varphi_{\boldsymbol{\nu}}(\boldsymbol{x}), \hat{f}_j(\pi_j(\boldsymbol{x})))| \\
&\leq \sum_{j=1}^{M} \sum_{\boldsymbol{\nu} \in I_j} \big(|\varphi_{\boldsymbol{\nu}}(\boldsymbol{x}) f_j(\pi_j(\boldsymbol{x})) - \varphi_{\boldsymbol{\nu}}(\boldsymbol{x}) f_j(\pi_j(\boldsymbol{x}))| \\
&\qquad + |\varphi_{\boldsymbol{\nu}}(\boldsymbol{x}) f_j(\pi_j(\boldsymbol{x})) - \Phi_\varepsilon^\times(\varphi_{\boldsymbol{\nu}}(\boldsymbol{x}), \hat{f}_j(\pi_j(\boldsymbol{x})))|\big) \\
&\leq \sup_{i \leq M} \|f_i - \hat{f}_i\|_{L^\infty(C_i)} \sum_{j=1}^{M} \sum_{\boldsymbol{\nu} \in I_j} |\varphi_{\boldsymbol{\nu}}(\boldsymbol{x})| + \sum_{j=1}^{M} \sum_{\{\boldsymbol{\nu} \in I_j \,|\, \boldsymbol{x} \in \operatorname{supp} \varphi_{\boldsymbol{\nu}}\}} \varepsilon \\
&\leq C N^{-\frac{k+s}{m}} + d\varepsilon \leq C N^{-\frac{k+s}{m}},
\end{aligned}$$

where we used that $\boldsymbol{x}$ can be in the support of at most d of the $\varphi_{\boldsymbol{\nu}}$, and where C is a constant depending on d and $\mathcal{M}$.

Finally, let us bound the size and depth of this approximation. Using $\operatorname{size}(\varphi_{\boldsymbol{\nu}}) \leq C$, $\operatorname{depth}(\varphi_{\boldsymbol{\nu}}) \leq C$ (see (5.3.9)) and $\operatorname{size}(\Phi_\varepsilon^\times) \leq C \log(\varepsilon) \leq C \log(N)$ and $\operatorname{depth}(\Phi_\varepsilon^\times) \leq C \operatorname{depth}(\varepsilon) \leq C \log(N)$ (see Lemma 7.3) we find

$$\begin{aligned}
\sum_{j=1}^{M} \sum_{\boldsymbol{\nu} \in I_j} \Big(\operatorname{size}(\Phi_\varepsilon^\times) + \operatorname{size}(\varphi_{\boldsymbol{\nu}}) + \operatorname{size}(\hat{f}_i \circ \pi_j)\Big) &\leq \sum_{j=1}^{M} \sum_{\boldsymbol{\nu} \in I_j} C \log(N) + C + C N \log(N) \\
&= O(N \log(N)),
\end{aligned}$$

which implies the bound on $\operatorname{size}(\Phi_N)$. Moreover,

$$\begin{aligned}
\operatorname{depth}(\Phi_N) &\leq \operatorname{depth}(\Phi_\varepsilon^\times) + \max\Big\{\operatorname{depth}(\varphi_{\boldsymbol{\nu}}, \hat{f}_j)\Big\} \\
&\leq C \log(N) + \log(N) = O(\log(N)).
\end{aligned}$$

This completes the proof. □

Bibliography and further reading

The ideas of Section 8.1 were originally developed in [10], with an extension to L^∞ approximation provided in [9]. These arguments can be extended to yield dimension-independent approximation rates for high-dimensional discontinuous functions, provided the discontinuity follows a Barron function, as shown in [173]. The Barron class has been generalized in various ways, as discussed in [136, 135, 238, 239, 11].

The compositionality assumption of Section 8.2 was discussed in the form presented in [176]. An alternative approach, known as the hierarchical composition/interaction model, was studied in [117].

The manifold assumption discussed in Section 8.3 is frequently found in the literature, with notable examples including [210, 39, 34, 202, 154, 116].

Another prominent direction, omitted in this chapter, pertains to scientific machine learning. High-dimensional functions often arise from (parametric) PDEs, which have a rich literature describing their properties and structure. Various results have shown that neural networks can leverage the inherent low-dimensionality known to exist in such problems. Efficient approximation of certain classes of high-dimensional (or even infinite-dimensional) analytic functions, ubiquitous in parametric PDEs, has been verified in [207, 208]. Further general analyses for high-dimensional parametric problems can be found in [165, 120], and results exploiting specific structural conditions of the underlying PDEs, e.g., in [123, 197]. Additionally, [57, 148, 164] provide results regarding fast convergence for certain smooth functions in potentially high but finite dimensions.

For high-dimensional PDEs, elliptic problems have been addressed in [77], linear and semilinear parabolic evolution equations have been explored in [78, 70, 99], and stochastic differential equations in [107, 79].

Exercises

Exercise 8.8. Let $C > 0$ and $d \in \mathbb{N}$. Show that, if $g \in \Gamma_C$, then

$$a^{-d} g\left(a(\cdot - \boldsymbol{b})\right) \in \Gamma_C,$$

for every $a \in \mathbb{R}_+$, $\boldsymbol{b} \in \mathbb{R}^d$.

Exercise 8.9. Let $C > 0$ and $d \in \mathbb{N}$. Show that, for $g_i \in \Gamma_C$, $i = 1, \dots, m$ and $c = (c_i)_{i=1}^m$ it holds that

$$\sum_{i=1}^m c_i g_i \in \Gamma_{\|c\|_1 C}.$$

Exercise 8.10. For every $d \in \mathbb{N}$ the function $f(\boldsymbol{x}) := \exp(-\|\boldsymbol{x}\|_2^2/2)$, $\boldsymbol{x} \in \mathbb{R}^d$, belongs to Γ_d. It holds $C_f = O(\sqrt{d})$, for $d \to \infty$.

Exercise 8.11. Let $d \in \mathbb{N}$, and let $f(\boldsymbol{x}) = \sum_{i=1}^\infty c_i \sigma_{\mathrm{ReLU}}(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle + b_i)$ for $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{a}_i\| = 1, |b_i| \leq 1$ for all $i \in \mathbb{N}$. Show that for every $N \in \mathbb{N}$, there exists a ReLU neural network with N neurons and one layer such that

$$\|f - f_N\|_{L^2(B_1^d)} \leq \frac{3\|c\|_1}{\sqrt{N}}.$$

Hence, every infinite ReLU neural network can be approximated at a rate $O(N^{1/2})$ by finite ReLU neural networks of width N.

Exercise 8.12. Let $C > 0$ prove that every $f \in \Gamma_C$ is continuously differentiable.