25

Expressiveness of Shallow Networks

This chapter is devoted to the approximation power of neural networks—their *expressive power*, in Deep Learning parlance. The first focal point is the famed *universal approximation theorem*, i.e., the fact that, for reasonable activation functions, every multivariate real-valued continuous function can be uniformly approximated on any compact set with arbitrary accuracy using functions that are generated by shallow networks. Next, concentrating on ReLU activation, the rate of approximation of Lipschitz functions by shallow networks is to be analyzed, leading to upper and lower estimates that almost match.

25.1 Activation Functions and Universal Approximation

In the univariate setting and with ReLU activation, the functions generated by shallow networks coincide with CPwL functions (see Theorem 24.1), so they are dense in any $C[a,b]$. This denseness result is to be extended to the multivariate setting and to other activation functions. In fact, the result below characterizes the activation functions for which denseness holds.

Theorem 25.1 *For a continuous activation function $\phi\colon \mathbb{R} \to \mathbb{R}$ and for a compact subset $\mathcal{X}$ of $\mathbb{R}^d$, let*

$$\mathcal{N}_\phi(\mathcal{X}) := \Big\{ g \in F(\mathcal{X}, \mathbb{R}) : \text{there are } n \geq 1,\ a_1, \dots, a_n \in \mathbb{R}^d, \text{ and } b, c \in \mathbb{R}^n \text{ such that } g(x) = \sum_{j=1}^{n} c_j \phi(\langle a_j, x \rangle + b_j) \text{ for all } x \in \mathcal{X} \Big\} \quad (25.1)$$

denote the set of functions generated by shallow networks with the activation function ϕ. The following properties are equivalent:

(i) the set $\mathcal{N}_\phi(\mathcal{X})$ is dense in $C(\mathcal{X})$;
(ii) the function ϕ is not a polynomial.

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

Proof of (i)⇒(ii) This implication is clear, because if ϕ was a polynomial, then the set $\mathcal{N}_\phi(\mathcal{X})$ would be contained in the space of polynomials of degree at most $\deg(\phi)$ and would therefore not be dense in $C(\mathcal{X})$. □

The core of the argument is the implication (ii)⇒(i). A first step consists in realizing that the problem can be reduced to the univariate case.

Proof of (ii)⇒(i)*, Step 1* Suppose that the denseness result holds in the case $d = 1$. Now, for $d > 1$, it is easy to verify that the set

$$\mathcal{A} = \operatorname{span}\left\{ f \in F(\mathcal{X}, \mathbb{R}) : f = \exp(\langle v, \cdot \rangle) \text{ for some } v \in \mathbb{R}^d \right\}$$

is a subalgebra of $C(\mathcal{X})$ that vanishes nowhere and separates points. Thus, by the *Stone–Weierstrass theorem* (Theorem E.3), it is dense in $C(\mathcal{X})$. Therefore, given a function $f \in C(\mathcal{X})$ and an accuracy $\varepsilon > 0$, one can find $k \geq 1$, $\gamma_1, \ldots, \gamma_k \in \mathbb{R}$, and $v_1, \ldots, v_k \in \mathbb{R}^d$ such that

$$\left| f(x) - \sum_{i=1}^{k} \gamma_i \exp(\langle v_i, x \rangle) \right| < \frac{\varepsilon}{2} \qquad \text{for all } x \in \mathcal{X}. \tag{25.2}$$

For each $i \in [1:k]$, the set $\{\exp(\langle v_i, x \rangle), x \in \mathcal{X}\}$ is a compact subset of $\mathbb{R}$, so by invoking the result for $d = 1$, one can find $n_i \geq 1$ and $a_i, b_i, c_i \in \mathbb{R}^{n_i}$ such that, for all $t \in \{\exp(\langle v_i, x \rangle), x \in \mathcal{X}\}$,

$$\left| \exp(t) - \sum_{j=1}^{n_i} c_{i,j} \phi(a_{i,j} t + b_{i,j}) \right| < \frac{\varepsilon}{2 \sum_{\ell=1}^{k} |\gamma_\ell|}. \tag{25.3}$$

One deduces from (25.2) and (25.3) that, for all $x \in \mathcal{X}$,

$$\begin{aligned}
&\left| f(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \gamma_i c_{i,j} \phi(a_{i,j} \langle v_i, x \rangle + b_{i,j}) \right| \\
&\leq \left| f(x) - \sum_{i=1}^{k} \gamma_i \exp(\langle v_i, x \rangle) \right| + \sum_{i=1}^{k} |\gamma_i| \left| \exp(\langle v_i, x \rangle) - \sum_{j=1}^{n_i} c_{i,j} \phi(a_{i,j} \langle v_i, x \rangle + b_{i,j}) \right| \\
&< \frac{\varepsilon}{2} + \sum_{i=1}^{k} |\gamma_i| \frac{\varepsilon}{2 \sum_{\ell=1}^{k} |\gamma_\ell|} = \varepsilon.
\end{aligned}$$

This means that f can be uniformly approximated by elements from $\mathcal{N}_\phi(\mathcal{X})$ with error at most ε. Since $f \in C(\mathcal{X})$ and $\varepsilon > 0$ were arbitrary, the denseness of $\mathcal{N}_\phi(\mathcal{X})$ in $C(\mathcal{X})$ is proved. □

The argument for the second step of the implication (ii)⇒(i) involves an identity known as the *Peano representation* of divided differences. Recall first

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

that the *divided difference* of a function f at points $t_0 < t_1 < \cdots < t_{k-1} < t_k$ is defined inductively by $[t_0]f = f(t_0)$ and, for $k \geq 1$, by

$$[t_0, t_1, \ldots, t_{k-1}, t_k]f = \frac{[t_1, \ldots, t_{k-1}, t_k]f - [t_0, t_1, \ldots, t_{k-1}]f}{t_k - t_0}.$$

For a k-times differentiable function f, the divided difference $[t_0, t_1, \ldots, t_k]f$ provides a numerical approximation to $f^{(k)}(x)$ when $t_0, t_1, \ldots, t_{k-1}, t_k$ are all close to x. In fact, the divided difference can be represented as

$$[t_0, t_1, \ldots, t_{k-1}, t_k]f = \frac{1}{k!} \int_{t_0}^{t_k} M_{t_0,\ldots,t_k}(t) f^{(k)}(t) dt \tag{25.4}$$

for some function $M_{t_0,\ldots,t_k}$ known as the L_1-normalized *B-spline* relative to $t_0, \ldots, t_k$. It is a piecewise polynomial of degree $< k$ with breakpoints $t_0, \ldots, t_k$, globally $(k-2)$-times continuously differentiable, nonnegative on its support $[t_0, t_k]$, and integrating to one. The identity (25.4) can be verified (readers are invited to do so in Exercise 25.1) by relying on the inductive definition of B-splines, which is given by $M_{t_0,t_1}(t) = \mathbb{1}_{[t_0,t_1]}(t)/(t_1 - t_0)$ and, for $k \geq 2$,

$$M_{t_0,\ldots,t_k}(t) = \frac{k}{k-1} \Big(\frac{t - t_0}{t_k - t_0} M_{t_0,\ldots,t_{k-1}}(t) + \frac{t_k - t}{t_k - t_0} M_{t_1,\ldots,t_k}(t) \Big). \tag{25.5}$$

Proof of (ii)⇒(i)*, Step 2* The objective is to establish the univariate result in the case $\phi \in C^\infty(\mathbb{R})$. Let a nonnegative integer k and a real number b be fixed for now. Given $x \in \mathcal{X}$, the *Peano representation* (25.4) for the divided differences at the points $0, h, \ldots, kh$ of the function $f_x \colon t \in \mathbb{R} \mapsto \phi(tx + b) \in \mathbb{R}$ is written as

$$[0, h, \ldots, kh]f_x = \frac{1}{k!} \int_0^{kh} M_{0,h,\ldots,kh}(t) x^k \phi^{(k)}(tx + b) dt.$$

Setting $\gamma := \max\{|u|, u \in \mathcal{X}\}$ and $\varepsilon_h := \max\{|\phi^{(k)}(v + b) - \phi^{(k)}(b)|, |v| \leq kh\gamma\}$, it follows that

$$\begin{aligned} \left| [0, h, \ldots, kh]f_x - \frac{\phi^{(k)}(b)}{k!} x^k \right| &= \left| \frac{x^k}{k!} \int_0^{kh} M_{0,h,\ldots,kh}(t) (\phi^{(k)}(tx + b) - \phi^{(k)}(b)) dt \right| \\ &\leq \frac{\gamma^k}{k!} \int_0^{kh} M_{0,h,\ldots,kh}(t) \varepsilon_h dt = \frac{\gamma^k}{k!} \varepsilon_h. \end{aligned}$$

Observing that the function $x \in \mathcal{X} \mapsto [0, h, \ldots, kh]f_x$ belongs to $\mathcal{N}_\phi(\mathcal{X})$ and that the bound $(\gamma^k/k!)\varepsilon_h$ tends to zero as $h \to 0$ independently of $x \in \mathcal{X}$, one deduces that the map $x \in \mathcal{X} \mapsto (\phi^{(k)}(b)/k!)x^k$ belongs to the closure $\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$ of $\mathcal{N}_\phi(\mathcal{X})$. Since there exists some $b \in \mathbb{R}$ such that $\phi^{(k)}(b) \neq 0$—otherwise ϕ would be a polynomial—one derives that the map $x \in \mathcal{X} \mapsto x^k$ itself belongs to $\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$. This being true for any integer $k \geq 0$, one concludes that

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

$\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$ contains all polynomials, and in turn, by the *Weierstrass theorem* (Theorem E.1), that $\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$ equals $C(\mathcal{X})$. □

The argument for the implication (ii)⇒(i) now requires a final step to remove the assumption that $\phi \in C^\infty(\mathbb{R})$. It consists in selecting a compactly supported function $\psi \in C^\infty(\mathbb{R})$ and in considering its *convolution product* with a merely continuous function $\phi\colon \mathbb{R} \to \mathbb{R}$. This convolution product is defined for any $x \in \mathbb{R}$ by

$$(\phi * \psi)(x) = \int_{-\infty}^{\infty} \phi(x-y)\psi(y)dy. \tag{25.6}$$

It can be verified that the function $\phi * \psi$ belongs to $C^\infty(\mathbb{R})$. The same holds when convolving with $\psi_\varepsilon\colon x \in \mathbb{R} \mapsto \psi(x/\varepsilon)/\varepsilon$ for any $\varepsilon > 0$. By choosing e.g. ψ to be the *bump function* $x \in \mathbb{R} \mapsto \mathbb{1}_{[-1,1]}(x) \times \exp(-1/(1-x^2))$ normalized so that $\int_{\mathbb{R}} \psi = 1$, there is the added bonus that $\phi * \psi_\varepsilon$ converges uniformly to ϕ on any compact subset of $\mathbb{R}$ when $\varepsilon \to 0$; see Exercise 25.2.

Proof of (ii)⇒(i)*, Step 3* For $\varepsilon > 0$, with the compactly supported function $\psi_\varepsilon \in C^\infty(\mathbb{R})$ chosen as above, one first observes that, for any $a, b \in \mathbb{R}$, the map

$$x \in \mathcal{X} \mapsto (\phi * \psi_\varepsilon)(ax+b) = \int_{-\infty}^{\infty} \phi(ax+b-y)\psi_\varepsilon(y)dy$$

belongs to $\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$. It follows that $\mathrm{cl}(\mathcal{N}_{\phi*\psi_\varepsilon}(\mathcal{X})) \subseteq \mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$. Assume now that $\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$ is a proper subset of $C(\mathcal{X})$. Invoking the Weierstrass theorem again, there exists an integer $k \geq 0$ such that the map $x \in \mathcal{X} \mapsto x^k$ does not belong to $\mathrm{cl}(\mathcal{N}_\phi(\mathcal{X}))$, and hence does not belong to $\mathrm{cl}(\mathcal{N}_{\phi*\psi_\varepsilon}(\mathcal{X}))$ either. However, according to Step 2, the map $x \in \mathcal{X} \mapsto ((\phi * \psi_\varepsilon)^{(k)}(b)/k!)x^k$ belongs to $\mathrm{cl}(\mathcal{N}_{\phi*\psi_\varepsilon}(\mathcal{X}))$ for any $b \in \mathbb{R}$. This implies that $(\phi*\psi_\varepsilon)^{(k)}(b) = 0$ for any $b \in \mathbb{R}$, i.e., that $\phi * \psi_\varepsilon$ is a polynomial of degree $< k$. It follows that ϕ, as the limit of $\phi * \psi_\varepsilon$ when $\varepsilon \to 0$, is also a polynomial of degree $< k$, which is not the case. This contradiction finishes the proof that $\mathcal{N}_\phi(\mathcal{X})$ is dense in $C(\mathcal{X})$. □

25.2 Approximation Rate with ReLU: Upper Bound

The universal approximation theorem (Theorem 25.1) is not quantitative: it says only that the error of best approximation to a given continuous function using shallow networks converges to zero as the width n goes to infinity, but it does not provide any information about the convergence speed. Concentrating on ReLU activation, one shall now target results about the approximation rate in terms of the number $(d+2)n \asymp dn$ of parameters describing the set of

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

d-variate functions generated by shallow ReLU networks of width n. In the same spirit as in Theorem 25.1, this set is written as

$$\mathcal{N}_{\mathrm{ReLU}}^n := \left\{ \sum\nolimits_{j=1}^n c_j \,\mathrm{ReLU}(\langle a_j, \cdot \rangle + b_j) : \, a_1, \ldots, a_n \in \mathbb{R}^d \text{ and } b, c \in \mathbb{R}^n \right\}$$

without including a final bias, since it can be obtained by choosing one of the a_i to be zero. The worst-case considerations below involve *Lipschitz functions*. Precisely, one defines a model set (already encountered in Chapter 11) by

$$\mathcal{K}_{\mathrm{Lip}} := \left\{ f \in C([0,1]^d) : |f|_{\mathrm{Lip}} := \sup_{x \neq x' \in [0,1]^d} \frac{|f(x) - f(x')|}{\|x - x'\|_\infty} \leq 1 \right\}.$$

The main result of this section consists of a nearly tight upper bound for the approximation rate of Lipschitz functions using shallow ReLU networks. The complete proof is omitted[1] and only the simple case of univariate functions is treated here.

Theorem 25.2 *There is a positive constant C_d such that, for any $n \geq 2$,*

$$\sup_{f \in \mathcal{K}_{\mathrm{Lip}}} \inf_{g \in \mathcal{N}_{\mathrm{ReLU}}^n} \|f - g\|_{C([0,1]^d)} \leq C_d \ln(n) \frac{1}{n^{1/d}}. \tag{25.7}$$

Sketch of proof when $d = 1$ Let a function $f \in C([0,1])$ satisfy $|f|_{\mathrm{Lip}} \leq 1$. For $n \geq 2$, consider the continuous piecewise linear function g with breakpoints at $x_0 = 0, \ldots, x_i = i/(n-1), \ldots, x_{n-1} = 1$ that interpolates the values $f(x_0), \ldots, f(x_i), \ldots, f(x_{n-1})$ there. As outlined in the proof of Theorem 24.1, this function can be generated by a shallow ReLU network of width n, i.e., $g \in \mathcal{N}_{\mathrm{ReLU}}^n$. Moreover, it also satisfies $|g|_{\mathrm{Lip}} \leq 1$, from where the inequality $\|f - g\|_{C([0,1])} \leq 1/(n-1)$ can be easily obtained (an improved inequality is provided in Lemma 26.5). Indeed, for any $x \in [0,1]$, choosing $i \in [0 : n-1]$ such that $|x - x_i| \leq 1/(2(n-1))$ leads to

$$\begin{aligned} |f(x) - g(x)| &\leq |f(x) - f(x_i)| + |g(x_i) - g(x)| \leq (|f|_{\mathrm{Lip}} + |g|_{\mathrm{Lip}}) |x - x_i| \\ &\leq \frac{1}{n-1}. \end{aligned}$$

In view of $n - 1 \geq n/2$, the bound $\inf \{\|f - g\|_{C([0,1])} : g \in \mathcal{N}_{\mathrm{ReLU}}^n\} \leq 2/n$ holds for any $f \in \mathcal{K}_{\mathrm{Lip}}$, meaning that the estimate (25.7) is valid when $d = 1$ even without the logarithmic factor. □

[1] The arguments are given in Bach (2017), with the result being stated in Subsection 4.7 there.

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

25.3 Approximation Rate with ReLU: Lower Bound

To justify the near-tightness of Theorem 25.2, this section provides a lower bound for the approximation rate of Lipschitz functions using shallow ReLU networks. Disregarding logarithmic factors, it matches the upper bound of the previous section.

Theorem 25.3 *There is a positive constant c_d such that, for any $n \geq 1$,*

$$\sup_{f \in \mathcal{K}_{\mathrm{Lip}}} \inf_{g \in \mathcal{N}^n_{\mathrm{ReLU}}} \|f - g\|_{C([0,1]^d)} \geq \frac{c_d}{\ln(2n)^{1/d}} \frac{1}{n^{1/d}}.$$

The result is a direct consequence of the following two observations, both of them being interesting in their own right.

Proposition 25.4 *Given any subset $\mathcal{G}$ of $C([0,1]^d)$, one has*

$$\sup_{f \in \mathcal{K}_{\mathrm{Lip}}} \inf_{g \in \mathcal{G}} \|f - g\|_{C([0,1]^d)} \geq \frac{1}{2\,\mathrm{vc}(\mathbb{1}_{(0,+\infty)} \circ \mathcal{G})^{1/d}},$$

where $\mathbb{1}_{(0,+\infty)} \circ \mathcal{G}$ denotes the family of boolean functions of the form $\mathbb{1}_{(0,+\infty)} \circ g$ for some $g \in \mathcal{G}$.

Proposition 25.5 *The set of shallow ReLU networks of width $n \geq 1$ yields a* VC-dimension *satisfying*

$$\mathrm{vc}(\mathbb{1}_{(0,+\infty)} \circ \mathcal{N}^n_{\mathrm{ReLU}}) \leq C d n \ln(2n)$$

for some absolute constant C that can be taken as $C = 40/\ln(2)$.

It now remains to justify these two propositions.

Proof of Proposition 25.4 The result is clear if $\delta \geq 1/2$, where

$$\delta := \sup_{f \in \mathcal{K}_{\mathrm{Lip}}} \inf_{g \in \mathcal{G}} \|f - g\|_{C([0,1]^d)}.$$

Thus, one assumes that $\delta < 1/2$ and considers the integer $n \geq 1$ such that $1/(2(n+1)) \leq \delta < 1/(2n)$. Let $\mathfrak{X} = \{x^{(\mathbf{i})} = [i_1/n; \ldots; i_d/n] : \mathbf{i} \in [0\!:\!n]^d\}$ be the set of $(n+1)^d$ nodes of the d-tensorized regular grid with spacing $1/n$. For each $\mathbf{i} \in [0\!:\!n]^d$, let $C_{\mathbf{i}}$ denote the cell associated with $x^{(\mathbf{i})}$ in the Voronoi tessellation of $[0,1]^d$ relative to the ℓ_∞-norm; see Figure 25.1. For any binary vector $\varepsilon \in \{0,1\}^{[0:n]^d}$, the function f defined for $x \in [0,1]^d$ by

$$f(x) = \sum_{\mathbf{i} \in [0:n]^d} \widetilde{\varepsilon}_{\mathbf{i}} \, \mathrm{dist}_{\ell_\infty}(x, [0,1]^d \setminus C_{\mathbf{i}}), \qquad \widetilde{\varepsilon}_{\mathbf{i}} := 2\varepsilon_{\mathbf{i}} - 1 \in \{-1, +1\}, \quad (25.8)$$

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

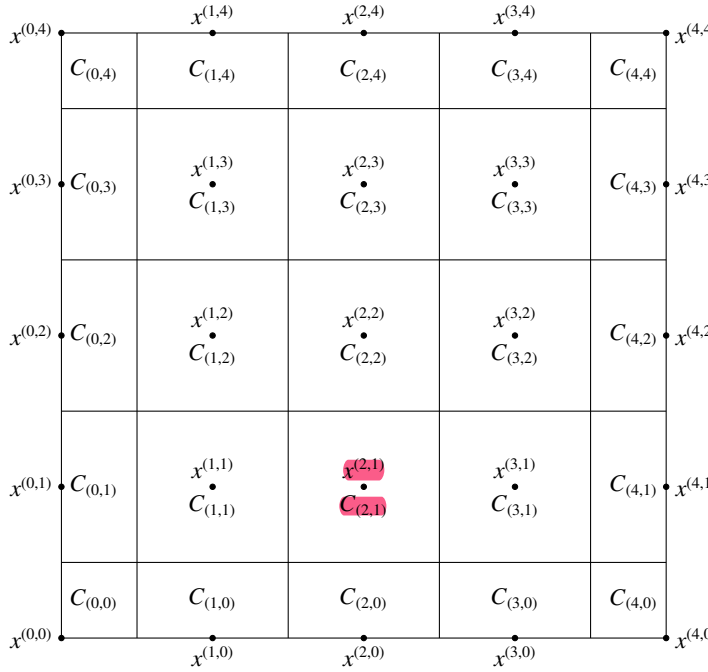

Figure 25.1 The cells $C_{\mathbf{i}}$ and their centers $x^{(\mathbf{i})}$ when $d = 2$ and $n = 4$.

can be verified to satisfy

$$|f|_{\mathrm{Lip}} \leq 1 \qquad \text{and} \qquad f(x^{(\mathbf{i})}) = \frac{\widetilde{\varepsilon}_{\mathbf{i}}}{2n} \quad \text{for all } \mathbf{i} \in [0:n^d].$$

Since there exists $g \in \mathcal{G}$ such that $|f(x) - g(x)| \leq \delta < 1/(2n)$ for all $x \in [0,1]^d$, one deduces that $\mathrm{sgn}(g(x^{(\mathbf{i})})) = \widetilde{\varepsilon}_{\mathbf{i}}$, i.e., that $(\mathbb{1}_{(0,+\infty)} \circ g)(x^{(\mathbf{i})}) = \varepsilon_{\mathbf{i}}$, for all $\mathbf{i} \in [0:n]^d$. This fact means that the set $\mathfrak{X}$ is shattered by $\mathbb{1}_{(0,+\infty)} \circ \mathcal{G}$. Therefore,

$$\mathrm{vc}(\mathbb{1}_{(0,+\infty)} \circ \mathcal{G}) \geq |\mathfrak{X}| = (n+1)^d \geq (1/(2\delta))^d,$$

which is a rearrangement of the announced result. □

Proof of Proposition 25.5 The main objective is to bound the *shatter function* (see Definition 2.1) of the family $\mathbb{1}_{(0,+\infty)} \circ \mathcal{N}^n_{\mathrm{ReLU}}$ as follows: for $m \geq (d+1)n$,

$$\begin{aligned} \tau(m) &:= \max_{x^{(1)},\dots,x^{(m)} \in \mathbb{R}^d} \left| \left\{ [\mathbb{1}_{(0,+\infty)}(h(x^{(1)})); \dots; \mathbb{1}_{(0,+\infty)}(h(x^{(m)}))], h \in \mathcal{N}^n_{\mathrm{ReLU}} \right\} \right| \\ &\leq \left(\frac{4m}{d\sqrt{n}} \right)^{4dn}. \end{aligned} \tag{25.9}$$

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

From here, with $\overline{m}$ denoting the VC-dimension of $\mathbb{1}_{(0,+\infty)} \circ \mathcal{N}^n_{\mathrm{ReLU}}$, one recalls that $\tau(\overline{m}) = 2^{\overline{m}}$. If $\overline{m} < (d+1)n$, then the result is immediately clear. If otherwise $\overline{m} \geq (d+1)n$, then the estimate (25.9) gives $2^{\overline{m}} \leq (4\overline{m}/(d\sqrt{n}))^{4dn}$. Taking the logarithm yields

$$\overline{m}\ln(2) \leq 4dn \ln\Big(\frac{4\overline{m}}{d\sqrt{n}}\Big), \quad \text{i.e.,} \quad \frac{4\overline{m}}{d\sqrt{n}} \leq \frac{16}{\ln(2)}\sqrt{n}\ln\Big(\frac{4\overline{m}}{d\sqrt{n}}\Big). \tag{25.10}$$

Since $\ln(t) < \sqrt{t}$ for any $t > 0$, this inequality implies

$$\frac{4\overline{m}}{d\sqrt{n}} \leq \frac{16}{\ln(2)}\sqrt{n}\sqrt{\frac{4\overline{m}}{d\sqrt{n}}}, \quad \text{and hence} \quad \frac{4\overline{m}}{d\sqrt{n}} \leq \frac{16^2}{\ln(2)^2} n \leq (2n)^{10}.$$

Substituting the latter into (25.10), one obtains the required estimate

$$\overline{m} \leq \frac{40}{\ln(2)} dn \ln(2n).$$

Turning to the justification of the bound (25.9), let $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^d$ be fixed from now on. By the positive homogeneity of ReLU, any $h \in \mathcal{N}^n_{\mathrm{ReLU}}$ can be written as $h = \sum_{j=1}^n \gamma_j \mathrm{ReLU}(\langle a_j, \cdot\rangle + b_j)$ where $a_1, \ldots, a_n \in \mathbb{R}^d$, $b_1, \ldots, b_n \in \mathbb{R}$, and importantly, $\gamma_1, \ldots, \gamma_n \in \{-1, +1\}$. Thus, the goal is to bound the cardinality of the set $\mathcal{S} \in \{0,1\}^m$ given by

$$\mathcal{S} := \bigcup_{\gamma_1,\ldots,\gamma_n \in \{-1,+1\}} \Big\{ \Big[\cdots ; \mathbb{1}_{(0,+\infty)}\Big(\sum_{j=1}^n \gamma_j \mathrm{ReLU}(\langle a_j, x^{(i)}\rangle + b_j)\Big); \cdots \Big] : a_1, \ldots, a_n \in \mathbb{R}^d, b_1, \ldots, b_n \in \mathbb{R} \Big\}. \tag{25.11}$$

For $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, notice that $\mathrm{ReLU}(\langle a, x^{(i)}\rangle + b)$ reduces to $\varepsilon_i(\langle a, x^{(i)}\rangle + b)$ with $\varepsilon_i := \mathbb{1}_{(0,+\infty)}(\langle a, x^{(i)}\rangle + b)$ for $i \in [1\!:\!m]$. The binary vector $\varepsilon \in \{0,1\}^m$ does not visit all 2^m possible configurations, though: it is restricted to a strict subset $\mathcal{E}$ of $\{0,1\}^m$. Indeed, since each $\varepsilon \in \mathcal{E}$ corresponds to an intersection-of-half-spaces region $\{[a;b] \in \mathbb{R}^{d+1} : \mathrm{sgn}(\langle a, x^{(i)}\rangle + b) = 2\varepsilon_i - 1,\ i \in [1\!:\!m]\}$ and since it is known (see Exercise 25.4 for the arguments) that the number of such intersection-of-half-spaces regions of $\mathbb{R}^k$ created by m hyperplanes is at most

$$R_{k,m} = 2\left[\binom{m-1}{0} + \binom{m-1}{1} + \cdots + \binom{m-1}{k-1}\right],$$

one obtains $|\mathcal{E}| \leq R_{d+1,m}$. Thus, for a particular choice of $\gamma_1, \ldots, \gamma_n \in \{-1, +1\}$,

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

the set appearing in the union (25.11) is included in

$$\bigcup_{\varepsilon^{(1)},\dots,\varepsilon^{(n)}\in\mathcal{E}} \Big\{\Big[\cdots;\mathbb{1}_{(0,+\infty)}\Big(\sum_{j=1}^{n}\gamma_j\varepsilon_i^{(j)}(\langle a_j,x^{(i)}\rangle+b_j)\Big);\cdots\Big]: a_1,\dots,a_n\in\mathbb{R}^d,b_1,\dots,b_n\in\mathbb{R}\Big\}. \quad (25.12)$$

For a particular choice of $\varepsilon^{(1)},\dots,\varepsilon^{(n)}\in\mathcal{E}$, the latter binary vectors correspond again to intersection-of-half-spaces regions created by m hyperplanes, but this time in the space $\mathbb{R}^{(d+1)n}$. Therefore, each set of binary vectors appearing in the union (25.12) has cardinality at most $R_{(d+1)n,m}$. All in all, the cardinality of the set (25.11) is bounded by $|\mathcal{S}|\le 2^n\times(R_{d+1,m})^n\times R_{(d+1)n,m}$. Invoking the estimate established in Lemma 2.6, it follows that

$$\begin{aligned}|\mathcal{S}| &\le 2^n\times 2^n\Big(\frac{e(m-1)}{d}\Big)^{dn}\times 2\Big(\frac{e(m-1)}{(d+1)n-1}\Big)^{(d+1)n-1}\\ &\le 2^n\times 2^n\Big(\frac{em}{d}\Big)^{dn}\times 2e^{-1}\Big(\frac{em}{dn}\Big)^{(d+1)n}\le 2^{2n}\Big(\frac{em}{d}\Big)^{2dn}\Big(\frac{em}{dn}\Big)^{2dn}\le\Big(\frac{4m}{d\sqrt{n}}\Big)^{4dn}.\end{aligned}$$

Since this is true for any choice of $x^{(1)},\dots,x^{(m)}\in\mathbb{R}^d$, the bound announced in (25.9) is now justified. □

Exercises

25.1 Verify that the function $M_{t_0,\dots,t_k}$ given by the inductive definition (25.5) is a piecewise polynomial of degree $< k$ with breakpoints $t_0,\dots,t_k$, is globally $(k-2)$-times continuously differentiable, is nonnegative on its support $[t_0,t_k]$, and integrates to one. Verify also the validity of Peano representation (25.4) of divided differences.

25.2 Show that the convolution product (25.6) of a compactly supported and infinitely differentiable function $\psi\in C^\infty(\mathbb{R})$ with a merely continuous function $\phi\in C(\mathbb{R})$ is infinitely differentiable, i.e., that $\phi*\psi\in C^\infty(\mathbb{R})$. Furthermore, if ψ is nonnegative, is supported on $[-1,1]$, and integrates to one, show that $|\phi(x)-(\phi*\psi_\varepsilon)(x)|\le\max\{|\phi(x)-\phi(x')|,|x-x'|\le\varepsilon\}$ for any $x\in\mathbb{R}$, where one defined $\psi_\varepsilon:=\psi(\cdot/\varepsilon)/\varepsilon$ for $\varepsilon>0$.

25.3 Fill in the details needed for a careful proof that the Lipschitz constant of the function f defined in (25.8) is at most one.

25.4 Let $R_{k,m}$, respectively $R_{k,m}^{\mathrm{aff}}$, denote the number of regions in $\mathbb{R}^k$ created by m hyperplanes, respectively affine hyperplanes, in general position.

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press

Prove by induction on $m \geq 1$ that

$$R_{k,m} = 2\left[\binom{m-1}{0} + \binom{m-1}{1} + \cdots + \binom{m-1}{k-1}\right],$$
$$R^{\text{aff}}_{k,m} = \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{k}.$$

To do so, assume without loss of generality that the $(m+1)$st (affine) hyperplane has equation $x_k = 0$ and count the number of regions added to the ones already created by the first m (affine) hyperplanes in order to obtain the recurrence relation

$$R^{(\text{aff})}_{k,m+1} = R^{(\text{aff})}_{k,m} + R^{(\text{aff})}_{k-1,m}.$$

https://doi.org/10.1017/9781009003933.037 Published online by Cambridge University Press