Recalling the definition of j^* and using the triangle inequality,

$$\begin{aligned}\left|E\langle\varepsilon,\hat{f}\rangle_n\right| &= \left|\frac{1}{n}E\sum_{i=1}^n \varepsilon_i(\hat{f}(X_i)-f_0(X_i))\right| \\ &\le \delta E\frac{1}{n}\sum_{i=1}^n|\varepsilon_i| + \left|\frac{1}{n}E\sum_{i=1}^n \varepsilon_i(f_{j^*}(X_i)-f_0(X_i))\right| \\ &\le \delta + \frac{1}{\sqrt{n}}E\left[|\xi_{j^*}|\|f_{j^*}-f_0\|_n\right].\end{aligned}$$

One further bounds $\|f_{j^*}-f_0\|_n \le \|f_{j^*}-\hat{f}\|_n + \|\hat{f}-f_0\|_n \le \delta + \|\hat{f}-f_0\|_n$ and via Cauchy-Schwarz,

$$\begin{aligned}E\left[|\xi_{j^*}|\|f_{j^*}-f_0\|_n\right] &\le \sqrt{2E\|\hat{f}-f_0\|_n^2+2\delta^2}\sqrt{E\left[\max_{1\le j\le\mathcal{N}_n}\xi_j^2\right]} \\ &\le \sqrt{2}\left[\sqrt{\hat{R}(\hat{f},f_0)}+\delta\right]\sqrt{3\log\mathcal{N}_n+1},\end{aligned}$$

where we have used the property on maxima mentioned above. Deduce

$$\left|E\langle\varepsilon,\hat{f}\rangle_n\right| \le \delta + \sqrt{2}\frac{\sqrt{4n}}{\sqrt{n}}\delta + \sqrt{2}\sqrt{\hat{R}(\hat{f},f_0)\cdot\frac{4\log\mathcal{N}_n}{n}} \le 5\delta + 4\sqrt{\hat{R}(\hat{f},f_0)\frac{\log\mathcal{N}_n}{n}},$$

where we use the assumption $1\le\log\mathcal{N}_n\le n$. One obtains

$$\hat{R}(\hat{f},f_0) \le R(f,f_0) + 4\sqrt{\hat{R}(\hat{f},f_0)\frac{\log\mathcal{N}_n}{n}} + 5\delta.$$

To conclude, one uses a similar argument as at the end of the proof of (4.16), so that one moves both $\hat{R}(\hat{f},f_0)$ terms to the left hand side of the inequality, which concludes the proof of (4.17).

Lemma 4.14. *Let $\xi_1,\dots,\xi_N$ be standard normal variables (but not necessarily independent). Then, for all $N\ge 1$,*

$$E\left[\max_{1\le j\le N}\xi_j^2\right] \le 3\log N + 1.$$

This is standard (see [SH20a], Lemma C.1, or e.g. [BLM13] Corollary 2.6 for a more general result for sub-exponential variables); the way to understand it: for Gaussian variables the maximum is of order at most $\sqrt{\log N}$ so the squares of N Gaussians have their maximum at most of size $(\log N)$.

4.4 Compositional structures: towards solving the curse of dimensionality

Discovering a hidden 'structure'. The 'raw' regression data collected by the statistician takes the form, in the setting model (4.1), of n vectors of size $d+1$: the n pairs (X_i^T, Y_i) with $X_i\in[0,1]^d$ and Y_i a real, with the dimension d possibly large (think for instance of e.g. $d=10$ or 20). The unknown regression function $f_0(x_1,\dots,x_d)$ depends on d of variables, and we have seen that if d is larger than a few units this may lead to a slow uniform convergence rate of the form $n^{-2\beta/(2\beta+d)}$ for the prediction risk. It is often the case though that the problem is effectively of smaller dimension than d. We give a number of frequently encountered examples

1. f_0 in fact depends on just one variable (but we do not know it a priori), for instance

$$f_0(x_1,\dots,x_d) = g(x_1),$$

for some $g : [0,1] \to \mathbb{R}$. In this case it seems reasonable to expect a rate $n^{-2\beta/(2\beta+1)}$, since the f_0 effectively depends on 1 variable only. More generally, f_0 may depend on a small number $t \le d$ of variables, although we do not know a priori which ones, e.g.

$$f_0(x_1,\dots,x_d) = g(x_2,x_3,x_d),$$

in which case the effective dimension should be 3, so we expect a rate $n^{-2\beta/(2\beta+3)}$.

2. In the preceding example, the function effectively depends on a small number of the original variables x_i, but it could depend on few variables only after transformation of the variables, for instance

$$f_0(x_1,\dots,x_d) = g(x_1+x_2+\cdots+x_d).$$

In this case $f_0(x_1,\dots,x_d) = g(x')$ only depends on 'one' variable $x' = x_1+\cdots+x_d$, so one expect a rate $n^{-2\beta/(2\beta+1)}$.

3. *Additive models.* It may be possible to write f_0 in an additive form

$$f_0(x_1,\dots,x_d) = \sum_{i=1}^{d} f_i(x_i),$$

for some functions $f_1,\dots,f_d$ depending on one variable only. If all functions f_i are at least β–Hölder, one expects a rate $d\cdot n^{-2\beta/(2\beta+1)}$ that is $n^{-2\beta/(2\beta+1)}$ if d is a fixed constant.

4. *Generalised additive models.* It may be possible to write f_0 in the form

$$f_0(x_1,\dots,x_d) = h\left(\sum_{i=1}^{d} f_i(x_i)\right),$$

for some real-valued functions $f_1,\dots,f_d$ (that are, as before, say all β–Hölder) and an unknown real 'link' function h that is γ–Hölder. One expects the rate to depend on β,γ, but not (too much) on the dimension d.

Class of compositions. In all the settings of the previous paragraph, one may note that the original function f_0 can be written as a composition of functions

$$f_0 = g_q \circ \cdots \circ g_1 \circ g_0,$$

for some integer $q \ge 1$. For instance, in the case of additive models one can set $g_0(x_1,\dots,x_d) = (f_1(x_1),\dots,f_d(x_d))$ (note that g_0 is then $\mathbb{R}^d$–valued) and $g_1(y_1,\dots,y_d) = y_1+\cdots+y_d$. For each of the examples in the above list, *if one knew beforehand* that f_0 is in one class of the other, one could certainly develop a specific estimation method using the special structure at hand. In practice, however, it would be desirable to have a method that is able to automatically 'learn the structure'. We are going to see that this is achieved by deep ReLU estimators.

Let us introduce the class, for $d = (d_0,\dots,d_{q+1}), t = (t_0,\dots,t_q), \beta = (\beta_0,\dots,\beta_q)$,

$$\mathcal{G}(q,d,t,\beta,K) = \Big\{ f = g_q\circ\cdots\circ g_0 : \quad g_i = (g_{ij})_j : [a_i,b_i]^{d_i} \to [a_{i+1},b_{i+1}]^{d_{i+1}}, \\ g_{ij} \in \mathcal{C}^{\beta_i}_{t_i}([a_i,b_i]^{t_i},K), \quad |a_i|,|b_i| \le K \Big\}, \tag{4.18}$$

where we denoted $\mathscr{C}_{t_i}^{\beta_i}$ for the Hölder ball over t_i variables to insist on the fact that these functions depend on t_i variables only (at most). The coefficients t_i can be interpreted as the maximal number of variables each function g_{ij} is allowed to depend on. In particular, this number is always at most d_i, but may actually be much smaller. Let us note that the decomposition of f_0 as a composition is typically not unique, but this is not of concern us here because we are interested in estimation of f_0 itself only.

Note that for $f_0 = g_1 \circ g_0$ with $d_1 = d_0 = t_1 = t_0 = 1$ and $\beta_0, \beta_1 \leq 1$, it follows from the definition of the Hölder class that f_0 has regularity $\beta_0\beta_1$, so that one expects a convergence rate of order $n^{-\frac{\beta_0\beta_1}{1+2\beta_0\beta_1}}$. It turns out that the actual (or 'effective') regularity depends on whether $\beta_i \leq 1$ or not. Let us define the following new 'regularity' parameter

$$\beta_i^* = \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1). \tag{4.19}$$

Convergence result for compositions. Given d, t, β as before, let us define the rate

$$\varepsilon_n^* = \max_{0\leq i\leq q} \left\{ n^{-\frac{\beta_i^*}{2\beta_i^*+t_i}} \right\}. \tag{4.20}$$

Example. For $d_0 = d_1 = t_0 = t_1 = q = 1$ and $f = g_1 \circ g_0$ with $\beta_1, \beta_0 \leq 1$, we have $\beta_0^* = \beta_0(\beta_1 \wedge 1) = \beta_0\beta_1$ and $\beta_1^* = \beta_1$, and the rate ε_n^* equals, since $\beta_0\beta_1 \leq \beta_1$,

$$\max\left(n^{-\frac{\beta_1}{2\beta_1+1}}, n^{-\frac{\beta_0\beta_1}{2\beta_0\beta_1+1}} \right) = n^{-\frac{\beta_0\beta_1}{2\beta_0\beta_1+1}},$$

which gives the rate announced above for this example. One may check that the formula (4.20) also gives the expected rate in the other examples above.

Theorem 4.15 (Convergence of ReLU DNNs for compositions)**.** *Suppose $f_0 \in \mathscr{G}(q, d, t, \beta, K)$ for arbitrary $\beta > 0$ and $K > 0$, integer q and vector of integers d, t. Let $\hat{f} = \hat{f}^{ReLU}$ be the estimator in (4.5) with $\mathscr{F} = \mathscr{F}(L, N, s, F)$ the class of realisations of neural networks with depth L, width vector $N = (N_l)_{1\leq l\leq L}$, sparsity s and uniform bound F. Suppose $F \geq K \vee 1$ and a choice of parameters, for ε_n^* as in (4.20),*

$$\log n \leq L \leq n\varepsilon_n^{*2}, \qquad n\varepsilon_n^{*2} \leq \min_{1\leq l\leq L} N_l \leq \max_{1\leq l\leq L} N_l \leq n^2, \qquad s \asymp (\log n)\left(n\varepsilon_n^{*2}\right).$$

Then there exists $C = C(q, d, t, \beta, F)$ such that

$$\sup_{f_0\in\mathscr{G}(q,d,t,\beta,K)} R(\hat{f}, f_0) \leq CL(\log n)^2\varepsilon_n^{*2}.$$

In particular, if $L \asymp \log n$, the maximum risk is bounded by $C(\log n)^3\varepsilon_n^{*2}$.

This result has a similar interpretation as Theorem 4.3: for well chosen parameters, the deep ReLU ERM achieves the rate ε_n^{*2} in prediction risk (up to a logarithmic factor). Moreover, this rate is optimal from the minimax perspective, as the next Theorem shows (under a mild condition on the dimensions). The remarkable point here is that, provided its parameters are well chosen, the deep ReLU estimator is able to *automatically* obtain the best possible rate, *without* being given any information beforehand on the underlying type of composition structure. Another setting somewhat different from compositional classes (but in the same spirit of a 'hidden structure') is that of data sitting on (or near) a geometric object, e.g. a manifold. One can show that deep ReLU ERM estimators again perform well in such settings, naturally 'adapting' to the unknown underlying geometric structure.

Theorem 4.16 (Minimax optimality for compositions)**.** *Consider the regression model (4.1), where the X_is are drawn from a distribution with density on $[0,1]^d$ which is bounded from above and below by positive constants. For arbitrary $\beta > 0$, integer q and vector of integers d, t, suppose $t_i \le \min(d_0, \dots, d_{i-1})$ for all i. Then for large enough K,*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{G}(q,d,t,\beta,K)} R(\hat{f}, f_0) \ge c\varepsilon_n^{*2},$$

where the infimum is taken over all possible estimators $\hat{f}$ of f in model (4.1).

Proof of Theorem 4.15. The proof is based again on the two key ingredients viewed in Section 4.2.3. These, combined with the Lemma below, enable to obtain the result. The proof is then very similar to that of Theorem 4.3. We sketch the proof now.

One first applies the oracle inequality Theorem 4.6. With the choice of parameters made in the statement of Theorem 4.15, and using the entropy control as in previous proofs, we directly see that the complexity term is of the expected order. So it is enough to focus on the approximation term, and derive an upper bound on the infimum of $\|f - f_0\|_\infty$ where f ranges in the class $\mathcal{F}$.

Step 1 (shift–and–rescale). One rewrites the composition $f_0 = g_q \circ \cdots \circ g_0$ as

$$f_0 = h_q \circ \cdots \circ h_0,$$

where now the function $h_{0j} \in \mathcal{C}_{t_0}^{\beta_0}([0,1]^{t_0}, 1)$, for $1 \le i \le q-1$ we have $h_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([0,1]^{t_i}, (2K)^{\beta_i})$ and $h_{qj} \in \mathcal{C}_{t_q}^{\beta_q}([0,1]^{t_q}, K(2K)^{\beta_q})$. This follows by shift–and–rescale by setting

$$h_0 = \frac{1}{2} + \frac{g_0}{2K}, \qquad h_i = \frac{1}{2} + \frac{g_i(2K\cdot - K)}{2K}, \qquad h_q = g_q(2K\cdot - K), \tag{4.21}$$

and checking that the one but last display holds. This is left as an exercise.

Step 2 (relating compositions).

Lemma 4.17. *Let $h_i = (h_{ij})$ be functions as in (4.21), with $K_i \ge 1$. Then for any functions $\tilde{h}_i = (\tilde{h}_{ij})$ with $\tilde{h}_{ij} : [0,1]^{t_i} \to [0,1]$, for and $C = C(K, \beta)$,*

$$\|h_q \circ \cdots h_0 - \tilde{h}_q \circ \cdots \circ \tilde{h}_0\|_{L^\infty[0,1]^d} \le C \sum_{i=0}^{q} \left\{ \left\| \left| h_i - \tilde{h}_i \right|_\infty \right\|_\infty \right\}^{\prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1)}.$$

Remark. Note that $h_i - \tilde{h}_i$ is a vector of functions: for each x one takes the maximum of the coordinates of the vector $(h_i - \tilde{h}_i)(x)$, and then the supremum norm $\left\| \left| h_i - \tilde{h}_i \right|_\infty \right\|_\infty$ is over $L^\infty[0,1]^{d_i}$.

Proof of Lemma 4.17. We set

$$H_i = h_i \circ \cdots \circ h_0, \quad \tilde{H}_i = \tilde{h}_i \circ \cdots \circ \tilde{h}_0.$$

By the triangle inequality, using that h_i is Q_i–Hölder for Q_i as specified above (4.21),

$$\begin{aligned} |H_i(x) - \tilde{H}_i(x)|_\infty &\le |h_i \circ H_{i-1}(x) - h_i \circ \tilde{H}_{i-1}(x)| + |h_i \circ H_{i-1}(x) - h_i \circ \tilde{H}_{i-1}(x)| \\ &\le Q_i \| |H_{i-1} - \tilde{H}_{i-1}|_\infty \|_\infty^{\beta_i \wedge 1} + \| |h_i - \tilde{h}_i|_\infty \|_\infty. \end{aligned}$$

Now we use this inequality recursively. To do so, first note that $(y+z)^\alpha \le y^\alpha + z^\alpha$ holds for $\alpha \in (0,1]$ and positive y, z (because the difference $(y+z)^\alpha - y^\alpha + z^\alpha$ is decreasing as a function of y). Noting

that all the powers $\beta_i \wedge 1$ are smaller than one, so that $Q_i^{\beta_i \wedge 1} \leq Q_i$ (using $Q_i \geq 1$ which follows from the assumption on K_i), a simple recursion gives, for any $1 \leq J \leq q$,

$$\|h_J \circ \cdots h_0 - \tilde{h}_J \circ \cdots \circ \tilde{h}_0\|_{L^\infty[0,1]^d} \leq \left(\prod_{i=1}^J Q_i\right) \sum_{i=0}^J \left\{\left\| |h_i - \tilde{h}_i|_\infty \right\|_\infty\right\}^{\prod_{\ell=i+1}^J (\beta_\ell \wedge 1)},$$

which gives the result by taking $J = q$. □

Step 3. Using Theorem 4.5, one can find functions $\tilde{h}_{ij}$ realisations of a ReLU network with $m \asymp \log n$, width of order $\mathcal{N}$ and sparsity $s_i \leq Cm\mathcal{N}$ with

$$\|\tilde{h}_{ij} - h_{ij}\|_\infty \lesssim \frac{\mathcal{N}}{n^2} + \mathcal{N}^{-\frac{\beta_i}{t_i}}.$$

One may assume $\tilde{h}_{ij}$ takes values in $[0,1]$ (up to applying ReLU and $\cdot \wedge 1$, which can be easily built with ReLU). Next one sets

$$f^* = \tilde{h}_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0.$$

Lemma 4.17 now gives

$$\|f^* - f_0\|_\infty \lesssim \sum_{i=0}^q \left(\mathcal{N}^{-\frac{\beta_i}{t_i}}\right)^{\prod_{\ell=i+1}^q (\beta_\ell \wedge 1)} \lesssim \max_{0 \leq i \leq q} \mathcal{N}^{-\frac{\beta_i^*}{t_i}} \lesssim \varepsilon_n^*,$$

as long as we set $\mathcal{N} := \lceil c \max_{0 \leq i \leq q} n^{\frac{t_i}{2\beta_i^* + t_i}} \rceil$ for $c > 0$ a small constant. To conclude, it is enough to update slightly f^* so that it takes values in $[-F, F]$ as requested. This is done through a simple shift–and–rescale argument and left to the reader. □

Proof of Theorem 4.16. We give the main idea of the proof in a simplified setting and briefly explain at the end how this extends to the general case.

Suppose that instead of the prediction loss, we want to prove the result in terms of the squared pointwise loss at $x = 0$, namely $\ell(f,g) = |f(0) - g(0)|^2$. We use Le Cam's two points argument (see a nonparametric statistics course): if one can find two functions f_{00} and f_{01} that both belong to the Hölder class over which the supremum is taken in the Theorem, that verify, for a small constant $c > 0$ (e.g. $c = 1/4$),

$$|f_{01}(0) - f_{00}(0)|^2 \gtrsim \varepsilon_n^{*2}$$
$$\mathrm{KL}(P_{f_{00}}, P_{f_{01}}) \lesssim c,$$

then the minimax rate in pointwise loss over the considered class of functions is bounded from below by a constant times $(\varepsilon_n^*)^2$.

Suppose for simplicity that all dimensions involved $t_i = d_i$ are equal to 1 and that the design points X_i are uniform over $[0,1]$. The latter implies for any f, g,

$$KL(P_f, P_g) = nE[(f(X_1) - g(X_1))^2] = n\|f - g\|_2^2.$$

Further define $i^* = \mathrm{argmin}_{i=0,\ldots,q} \beta_i^*/(2\beta_i^* + 1)$ as an index for which the estimation rate as in the statement of the Theorem is obtained.

Let us set $g_\ell(x) = x$ for $\ell < i^*$, $g_\ell(x) = x^{1 \wedge \beta_\ell}$ for $\ell > i^*$ and $g_{i^*}(x)$ to be chosen below, and

$$f_0(x) = g_q \circ \cdots \circ g_1 \circ g_0(x) = \left(g_{i^*}(x)\right)^{\prod_{\ell=i^*+1}^q \beta_\ell \wedge 1}. \tag{4.22}$$

Consider a smooth kernel function K and set $\tilde{g}(x) = h^{\beta_{i^*}} K(x/h)$. Under standard assumptions on the kernel, $\tilde{g}$ is β_{i^*}–Hölder. Now define two functions f_{01}, f_{00}, that are of the type of f_0 in (4.22), with respectively $g_{i^*}(x) = 0$ and $g_{i^*}(x) = \tilde{g}(x)$, that is

$$f_{00}(x) = 0, \qquad f_{01}(x) = \left(h^{\beta_{i^*}} K(x/h)\right)^{\prod_{\ell=i^*+1}^{q} \beta_\ell \wedge 1}.$$

Assuming $K(0) > 0$ one gets $|f_{00}(0) - f_{01}(0)| \gtrsim h^{\beta^*_{i^*}}$. Then evaluating the L^2–distance between $f_{00}(0)$ and f_{01} and using the previous identity for the KL divergence under Gaussian errors gives $\mathrm{KL}(P_{f_{00}}, P_{f_{01}}) \lesssim nh^{2\beta^*_{i^*}+1}$, which is a small positive constant if one choses $h \asymp n^{\frac{1}{2\beta^*_{i^*}+1}}$. Using the two-points lower bound scheme as mentioned above, one obtains the result for the squared pointwise loss.

For the prediction loss, the proof uses the same ideas, but the technique is slightly different: one uses (for instance) a 'many hypotheses' lower bound technique (again, see a nonparametric statistics course). One then replaces the perturbation constructed above around point $x = 0$ by 'many' perturbations around different points of $[0,1]$. The number of perturbations is controlled using Varshamov–Gilbert's lemma. This part of the proof is analogous to the proof of the minimax lower bound in one dimension in terms of the L^2–risk, so details are omitted. □

4.5 Minimax optimality and link to approximability

We conclude this chapter by asking a few questions the attentive reader may have already at the light of the results of the present and previous chapters

a) the upper-bound results we have obtained so far for regression consider ReLU DNNs with a logarithmic depth $L \asymp \log n$. On the other hand, results in Chapter 2 indicate that there exist deep networks that admit faster approximation properties in terms of the sparsity s, at the cost of choosing discontinuous weight functions. Can one make a link between both types of results?

b) from the remark in the previous point it is tempting to think that, since there exist faster approximation rates, it is perhaps possible to improve upon the convergence rates for deep ReLU estimators by taking a possibly different architecture in the network (e.g. by taking a deeper but thinner network). On the other hand, Theorem 4.16 provides a minimax lower bound, which matches the already obtained upper-bounds with ReLU networks of logarithmic depths (up to log terms)...

c) more generally, is there a link between approximability results – say of the type that given a network with some architecture, one cannot do better than a certain rate depending of the parameters of the architecture to approximate (say) Hölder functions – and minimax-type results – which assert that the estimation rate for estimating Hölder functions cannot be better than the minimax rate–?

To give a (partial) answer to a)–b) above, let us for simplicity restrict to a class of bounded and Lipschitz functions. Then one may use neural networks with the architecture considered in Chapter 2, Theorem 2.7.

Lemma 4.18 (estimating Lipschitz functions with polynomial depth)**.** *Suppose the true unknown $f_0 \in \mathscr{C}^1([0,1]^d, K)$ for some $K > 0$. Let $\hat{f} = \hat{f}^{ReLU}$ be the estimator in (4.5) with $\mathscr{F} = \mathscr{F}(L, N, s, F)$. Suppose $F \geq K \vee 1$ and a choice of parameters as follows, for $\beta = 1$,*

$$L \asymp \left(\frac{n}{\log n}\right)^{\frac{d}{2\beta+d}}, \qquad \min_{1\leq l\leq L} N_l \asymp \max_{1\leq l\leq L} N_l \asymp 1,$$

and sparsity $s \asymp L$. Then there exists $C = C(d, F)$ such that

$$\sup_{f_0 \in \mathscr{C}^1([0,1]^d, K)} R(\hat{f}, f_0) \leq C\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+d}}.$$

As is apparent from the proof of the Lemma below, even though for very deep ReLU–networks with constant width the approximation of Lipschitz functions is faster than for architectures with logarithmic depth and polynomial (in n) width, the global estimation rate is not significantly faster (there is a slight improvement in the logarithmic factor). This is of course expected, as one cannot go (uniformly) below the minimax rate! What happens is that the 'complexity' term arising in the study of $\hat{f} = \hat{f}^{ReLU}$ (the oracle inequality part) is much larger for the former architecture.

Proof. The oracle inequality Theorem 4.6 gives, for $\delta = 1/n$,

$$R(\hat{f}, f_0) = E\left[(\hat{f}(T) - f_0(T))^2\right] \lesssim \inf_{f\in\mathscr{F}} \|f - f_0\|_\infty^2 + \frac{\log \mathscr{N}(n^{-1}, \mathscr{F}, \|\cdot\|_\infty)}{n}.$$

Now combining with Theorem 2.7 and Lemma 4.13 to control the entropy term one obtains

$$R(\hat{f}, f_0) \lesssim s^{-\frac{4}{d}} + \frac{s \log(2LV^2 n)}{n} \lesssim s^{-\frac{4}{d}} + \frac{sL \log n}{n},$$

recalling that $V \leq (2n^2)^L$. Since $s \asymp L$, one can write $s^2 \asymp sL$ an write the previous upper bounds in terms of the quantity sL only as

$$R(\hat{f}, f_0) \lesssim (sL)^{-\frac{2}{d}} + \frac{(sL)\log n}{n} \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{2+d}},$$

using as usual that $x \to x^{-2/d} + a_n x$ is minimal for x such that two terms in the sum are of the same order, that is $x^{1+2/d} \asymp a_n$, which gives the announced result. □

Theorem 4.19. *Let $\beta, K > 0$ and $d \geq 1$ an integer. There exists a constant $c_1 > 0$ such that for $s \wedge L \geq c_1$,*

$$\sup_{f_0 \in \mathscr{C}^\beta([0,1]^d, K)} \inf_{f \in \mathscr{F}(L,N,s)} \|f - f_0\|_\infty \gtrsim \left(\frac{\log^{-1}(sL)}{sL}\right)^{\frac{\beta}{d}}.$$

Theorem 4.19 asserts that given a network architecture with depth L and sparsity s, one cannot achieve better than the rate $(sL)^{-\beta/d}$ to approximate β–Hölder functions in dimension d (up to a log term). As its proof reveals, it is a fairly direct consequence of combining the minimax lower bound in Theorem 4.16 in the simple case where $q = 1$ (i.e. the usual nonparametric rate for β–Hölder functions in dimension d) with the oracle inequality Theorem 4.6.

This result extends Proposition 2.6, which applies to Lipschitz functions, to any smoothness level $\beta > 0$. Results in this Chapter show that the lower bound of Theorem 4.19 is attained (up to log factors) for ReLU networks of logarithmic depth, as then $sL \asymp s(\log n) \asymp (\log n) n^{d/(2\beta+d)}$ and it suffices to apply Theorem 4.5. Results in Chapter 2 show that for $\beta = 1$ the lower bound of Theorem 4.19 is attained (up to log factors) for ReLU networks of depth polynomial in n and of constant width, as then $sL \asymp s^2$, which gives the quadratic dependence in s observed in Chapter 2.

Proof. For simplicity let us denote $\mathscr{C} = \mathscr{C}^\beta([0,1]^d, K)$ and $\mathscr{F} = \mathscr{F}(L, N, s)$. The oracle inequality Theorem 4.6 gives, taking the supremum over f_0s, and bounding the prediction risk in terms of the supremum norm, for n a large enough integer,

$$\begin{aligned}\sup_{f_0 \in \mathscr{C}} R(\hat{f}, f_0) &\lesssim \sup_{f_0 \in \mathscr{C}} \inf_{f \in \mathscr{F}} \|f - f_0\|_\infty^2 + \frac{\log \mathscr{N}(n^{-1}, \mathscr{F}, \|\cdot\|_\infty)}{n} \\ &\lesssim \sup_{f_0 \in \mathscr{C}} \inf_{f \in \mathscr{F}} \|f - f_0\|_\infty^2 + \frac{sL \log(sn)}{n},\end{aligned}$$

where we have used the entropy bound as in previous proofs. On the other hand, the maximum risk in the previous display can itself be bounded from below by

$$\sup_{f_0 \in \mathscr{C}} R(\hat{f}, f_0) \geq \inf_{\tilde{f}} \sup_{f_0 \in \mathscr{C}} R(\tilde{f}, f_0) \gtrsim n^{-\frac{2\beta}{2\beta+d}}.$$

Putting together the two previous bounds leads to

$$\sup_{f_0 \in \mathscr{C}} \inf_{f \in \mathscr{F}} \|f - f_0\|_\infty^2 + \frac{sL \log(sn)}{n} \gtrsim n^{-\frac{2\beta}{2\beta+d}}.$$

Now let us optimise with respect to the integer n, recalling that the above holds for any n large enough. The term $n^{-\frac{2\beta}{2\beta+d}}$ is larger than a large enough constant times $sL\log(sn)/n$ if $n^{\frac{d}{2\beta+d}} \asymp sL\log(sn)$. For such n one has $sL\log(sn) \asymp sL\log(sL)$. Since then $n^{-\frac{2\beta}{2\beta+d}} \asymp (sL\log(sL))^{-2\beta/d}$ the result follows. □

Bibliography

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[SH20a] Johannes Schmidt-Hieber. Appendix to “nonparametric regression using deep neural networks with relu activation function". *The Annals of Statistics*, 2020.

[SH20b] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.