Chapter 3

Complexity of neural networks

Contents

In this chapter, we evaluate the complexity of a class of ReLU neural networks. Section 3.1 presents an upper bound on the Vapnik–Chervonenkis (VC) dimension of such a class, when the support is fixed (that is, when the locations of the non-zero weights are fixed). Section 3.2 provides an upper bound on the covering number and entropy of such a class, when the sparsity is fixed (that is, when the number of non-zero weights are fixed, while their locations are left free). Both quantities are useful to evaluate the variance term of a statistical risk in different contexts. As we will see, these terms increase with the depth of the NN in a specific manner.

3.1 VC dimension of a class of neural networks

This section is mostly based on the lecture notes [Ger21], themself based on the paper [BHLM19].

In supervised classification, the VC dimension of a family $\mathscr{H}$ of classifiers is a key tool to evaluate the stochastic error. We start by recalling some well known facts.

3.1.1 VC dimension

We recall here some basic notions on the VC dimension, see, e.g., [MRT18]. Let us consider the problem of classifying points $x \in \mathbb{R}^d$ as label -1 or 1. Classifiers are measurable functions from $\mathbb{R}^d$ to $\{-1,1\}$. For some set of classifiers $\mathscr{H}$ let us introduce the shattering coefficient: for all $m \geq 1$,

$$\mathbb{S}_{\mathscr{H}}(m) = \max_{x_1,\dots,x_m \in \mathbb{R}^d} \#\{(h(x_1),\dots,h(x_m)),\ h \in \mathscr{H}\}, \tag{3.1}$$

which corresponds to the maximum number of different labeling for m points that $\mathscr{H}$ can produce. For instance, if they can be arbitrarily labeled, then $\mathbb{S}_{\mathscr{H}}(m) = 2^m$. But if the class $\mathscr{H}$ is small, we can have $\mathbb{S}_{\mathscr{H}}(m) < 2^m$.

As an illustration, let us consider the class of affine classifiers defined by

$$\mathscr{H} = \{x \in \mathbb{R}^d \mapsto \text{sign}(a^T x + b),\ a \in \mathbb{R}^d, b \in \mathbb{R}^d\}.$$

Then we easily check that when $d = 2$, we have $\mathbb{S}_{\mathscr{H}}(2) = 2^2$, $\mathbb{S}_{\mathscr{H}}(3) = 2^3$, $\mathbb{S}_{\mathscr{H}}(4) = 14 < 2^4$.

Next, the Vapnik–Chervonenkis (VC) dimension of a classifier set $\mathscr{H}$ is defined as the maximum number of points that can be arbitrary labelled with $\mathscr{H}$, that is,

$$d_{\text{VC}}(\mathscr{H}) = \sup\{m \geq 0 : \mathbb{S}_{\mathscr{H}}(m) = 2^m\} \tag{3.2}$$

(convention $\mathbb{S}_{\mathscr{H}}(0) = 1$). By definition of the shattering coefficient, $d_{\text{VC}}(\mathscr{H})$ corresponds to the maximum number m such that there exist $x_1, \dots, x_m \in \mathbb{R}^d$ with $\#\{(h(x_1), \dots, h(x_m)) = 2^m$. For instance, in the example above, the set of affine classifiers has a VC dimension equals to 3 in dimension 2. More generally, we can prove that it is $d+1$ in dimension $d \geq 1$ (see Exercise 9.5.2 in [Gir15]).

As an illustration, the shattering number/VC dimension is useful to bound the stochastic error for the empirical risk minimizer. For instance, the following result is classical.

Theorem 3.1. *[Theorem 9.1 and Corollary 9.7 in [Gir15]] Let (X_i, Y_i), $1 \leq i \leq n$, be i.i.d. copies of $(X, Y) \in \mathbb{R}^d \times \{-1, 1\}$ some random variables. Consider any set of classifiers $\mathscr{H}$ (measurable functions from $\mathbb{R}^d$ to $\{-1, 1\}$) and let $L(h) = P(Y \neq h(X))$ for all $h \in \mathscr{H}$. Then the empirical risk minimizer $\hat{h}_n \in \text{argmin}_{h \in \mathscr{H}} \sum_{i=1}^n \mathbb{1}\{h(X_i) \neq Y_i\}$ is such that*

$$L(\hat{h}_n) - \min_{h \in \mathscr{H}} L(h) \leq 2\sqrt{\frac{2\log(2\mathbb{S}_{\mathscr{H}}(n))}{n}} + \sqrt{\frac{t}{2n}}$$

$$L(\hat{h}_n) - \min_{h \in \mathscr{H}} L(h) \leq 4\sqrt{\frac{2d_{\text{VC}}(\mathscr{H})\log(2n+2)}{n}} + \sqrt{\frac{2t}{n}},$$

each of these inequalities holding with probability at least $1 - e^{-t}$.

3.1.2 Result

Here, we are going to upper-bound the VC dimension of the family of classifier corresponding to realizations of a NN *with a fixed support* (see definition in Chapter 1). Also, for simplicity, we will focus on the ReLU activation function. This means that we consider the set of NN classifiers

$$\mathscr{H} = \{h_w = \text{sign} \circ R(\Phi_w),\ w \in \mathbb{R}^s\}, \tag{3.3}$$

where Φ_w is the ReLU NN with input dimension d, output dimension 1, and the following fixed topology:

- depth $L \geq 2$;
- for each layer ℓ, a width $N_\ell \geq 1$, and an overall width $N = \sum_{\ell=1}^L N_\ell \geq 3$;
- for each layer ℓ, a sparsity $s_\ell \geq 0$, and an overall sparsity $s = \sum_{\ell=1}^L s_\ell \geq N$;
- support sets with according sizes which will contain the supports of the network;
- weight values in these support sets gathered in the vector $w \in \mathbb{R}^s$;

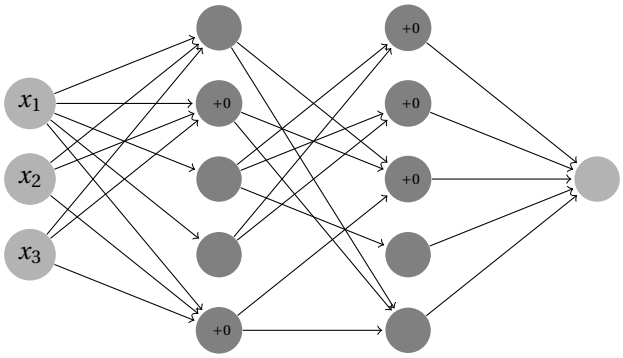

Figure 3.1: An instance of NN topology with $L = 3$, $N = 14$, $s_1 = 14$, $s_2 = 13$, $s_3 = 6$ and $s = 33$ which generates a classifier family $\mathscr{H}$ as in (3.3) for $d = 3$ and indexed by $w \in \mathbb{R}^{33}$.

Note that since w is allowed to contain zero values, s_ℓ is in fine only an upper bound on the sparsity of the ℓ-th layer of the network Φ_w.

In Figure 3.1, we have displayed an instance of such a possible topology, where the arrows code for the positions of 25 active weights (the support is fixed here). Plus the $5+5+1 = 11$ constant weights, this gives $s = 36$ and thus $w \in \mathbb{R}^{36}$ in (3.3).

Also, it is worth to note that, when exploring $\mathscr{H}$, the moving variable is w. (Note that the variable is not the input of the network here!). It is easy to see that $R(\Phi_w)$ is thus a piecewise s-multivariate polynomial (with a degree at most L). Hence, the maximal number of different labeling of m points that classifiers in $\mathscr{H}$ can produce is limited by this constraint. This observation is the basis to show the next result.

Theorem 3.2. *[Theorem 7 in [BHLM19]] Consider the set $\mathscr{H}$ of NN with the above topology. Then for all integer $m \geq s$, we have $\mathbb{S}_{\mathscr{H}}(m) \leq (4emL)^{sL}$ and $d_{\text{VC}}(\mathscr{H}) \leq 6sL\log_2(4eN)$.*

Note that this bound is almost sharp, in the sense that there exists a constant $c > 0$, such that for each $s \geq cL$ and $L \geq c$, there exists a ReLU NN of depth L and sparsity s such that the corresponding classifier class satisfies $d_{\text{VC}}(\mathscr{H}) \geq sL\log(s/L)/c$, see Theorem 3 in [BHLM19].

Theorem 3.2 can be extended to more general activation function: it holds for any activation function which is piece-wise polynomial with at most $p+1$ pieces and with degree bounded by $r \geq 0$.

- For $r = 0$ (piece-wise constant activation function), the bound becomes

$$d_{\text{VC}}(\mathscr{H}) \leq L + s\log_2(4epN\log_2(2epN)), \tag{3.4}$$

  which is of order $s\log(ps)$. In particular, this yields the desired bound (2.2) in Chapter 2.

- For $r \geq 1$, the bound becomes

$$d_{\text{VC}}(\mathscr{H}) \leq L + sL\log_2(4epR\log_2(2epR)), \tag{3.5}$$

  with $R = N + N(L-1)r^{L-1}$. This bound is of order $Ls\log(pN) + L^2 s\log(r)$.

Combined with Theorem 3.1, the above result gives an upper bound on the stochastic error term when optimizing a NN by minimizing the misclassification rate. However, note that this optimization is made with a fixed support of the network, which does not correspond to what is done in practice. Furthermore, to make a proper choice of the parameter s, N, L this stochastic error term should be balanced with the approximation error (bias term), which is not studied here (recall that some approximation error term are computed in Chapter 2). A more complete study addressing these points can be found in [BSH21] in the classification context, and in Chapter 4 in the regression context.

3.1.3 Proof of Theorem 3.2

Let us first prove the bound on the shattering number. More precisely, we establish

$$\mathbb{S}_{\mathscr{H}}(m) \le (4emL)^{s\bar{L}},$$

where $\bar{L} = s^{-1}\sum_{i=1}^{L} \bar{s}_i \in [1, L]$ with $\bar{s}_i = \sum_{\ell=1}^{i} s_\ell$ denoting the number of weights of layers between 1 and i.

Let $x_1, \dots, x_m \in \mathbb{R}^d$. The aim is to bound by above

$$\#\{(\mathrm{sign}(h_w(x_1)), \dots, \mathrm{sign}(h_w(x_m)),\ w \in \mathbb{R}^s\}$$

the number of different sign vectors that we can form by using $\mathscr{H}$. The idea is to build a partition $\{P_i\}_{1\le i\le M}$ of $\mathbb{R}^s$, of size $M \ge 1$, such that for all i, j, the function $w \in P_i \mapsto h_w(x_j)$ coincides with a s-multivariate polynomial with degree at most L. Then, we have

$$\begin{aligned}\#\{(\mathrm{sign}(h_w(x_1)), \dots, \mathrm{sign}(h_w(x_m)),\ w \in \mathbb{R}^s\} &\le \sum_{i=1}^{M} \#\{(\mathrm{sign}(h_w(x_1)), \dots, \mathrm{sign}(h_w(x_m)),\ w \in P_i\} \\ &\le \sum_{i=1}^{M} \#\{(\mathrm{sign}(p_1(w)), \dots, \mathrm{sign}(p_m(w))),\ w \in \mathbb{R}^s\},\end{aligned}$$

where each p_j is a s-multivariate polynomial with degree at most L on P_i. We can now apply the following lemma.

Lemma 3.3 (Theorem 8.3 in [ABB+99])**.** *For any polynomial $p_1, \dots, p_m$ with degree at most $d \ge 1$ in $n \le m$ variables, we have*

$$\#\{(sign(p_1(w)), \dots, sign(p_m(w))),\ w \in \mathbb{R}^n\} \le 2(2emd/n)^n.$$

Hence, Lemma 3.3 gives the bound

$$\mathbb{S}_{\mathscr{H}}(m) \le M2(2emL/s)^s,$$

provided that we have built a partition $\{P_i\}_{1\le i\le M}$ of $\mathbb{R}^s$ as above. In Section 3.1.4, we are going to prove that this is possible with M smaller than a given bound. More precisely, we are going to prove the following property.

Property 3.4. *For all $x_1, \dots, x_m \in \mathbb{R}^d$, there is a partition $\{P_i\}_{1\le i\le M}$ of $\mathbb{R}^s$, of size $M \ge 1$, such that for all i, j, the function $w \in P_i \mapsto h_w(x_j)$ coincide with a s-multivariate polynomial with degree at most L and with M satisfying*

$$M \le \prod_{\ell=1}^{L-1} 2(2em\ell N_\ell/\bar{s}_\ell)^{\bar{s}_\ell}. \tag{3.6}$$

Then, (3.6) and the above display entails (with $N_L = 1$)

$$\mathbb{S}_{\mathscr{H}}(m) \le \prod_{\ell=1}^{L} 2(2em\ell N_\ell/\bar{s}_\ell)^{\bar{s}_\ell} \le 2^L \left(\frac{\sum_{\ell=1}^{L} 2em\ell N_\ell}{\sum_{\ell=1}^{L} \bar{s}_\ell}\right)^{\sum_{\ell=1}^{L}\bar{s}_\ell}.$$

The last inequality holds because the geometrical average is smaller than the arithmetic average: for all $y_1, \dots, y_L > 0$, $a_1, \dots, a_L \ge 0$ with $\sum_{\ell=1}^{L} a_\ell > 0$, we have

$$\left(\prod_{\ell=1}^{L} y_\ell^{a_\ell}\right)^{1/\sum_{\ell=1}^{L} a_\ell} \le \frac{\sum_{\ell=1}^{L} a_\ell y_\ell}{\sum_{\ell=1}^{L} a_\ell}.$$

(Apply this with $a_\ell = \bar{s}_\ell$ and $y_\ell = 2emN_\ell\ell/\bar{s}_\ell$). Now, since $\sum_{\ell=1}^{L} \ell N_\ell \le NL$, $s\bar{L} = \sum_{\ell=1}^{L} \bar{s}_\ell$ and $\max(L,N) \le s \le s\bar{L}$, we obtain

$$\mathbb{S}_{\mathscr{H}}(m) \le 2^L \left(\frac{2emNL}{\sum_{\ell=1}^{L} \bar{s}_\ell} \right)^{s\bar{L}} \le (4emL)^{s\bar{L}},$$

which is the desired bound because $\bar{L} \le L$.

Finally, by applying Lemma 3.5 (used with $r = 2eNL$, $t = L$, $w = s\bar{L}$), we deduce

$$d_{\text{VC}}(\mathscr{H}) \le L + s\bar{L}\log_2(4eNL\log_2(2eNL)) \le sL(1 + s\log_2(4eNL\log_2(2eNL))).$$

Since $1 + s\log_2(4eNL\log_2(2eNL)) \le 3s\log_2(4eNL) \le 6s\log_2(4eN)$, because $L \le N$. This provides the desired bound on the VC dimension.

Lemma 3.5. *For all $r \ge 16$, and $w \ge t > 0$, for all $m \ge t + w\log_2(2r\log_2 r)$, we have $2^m > 2^t(mr/w)^w$.*

The proof of Lemma 3.5 is a direct consequence of the fact that $\Psi : x \mapsto x - t - w\log_1(xr/w)$ is nondecreasing on $[w\log 2, +\infty)$ with a non-negative value in $x_0 = t + w\log_2(2r\log_2 r)$. Indeed, $\Psi(x_0) \ge w\log_2\left(\frac{2\log_2 r}{1+\log_2(2r\log_2 r)}\right)$, which is a non-decreasing value of $r \ge 16$, with value 0 in $r = 16$.

3.1.4 Proving Property 3.4

For simplicity of the exposition, we assume here that the input dimension is $d = 1$ (but once you have read this with $d = 1$, the case of a general d is totally analogue).

Let us fix any $x_1, \ldots, x_m \in \mathbb{R}^d$. Let us build a sequence of $\mathbb{R}^s$-partitions[1] $\mathscr{S}_0 = \{\mathbb{R}^s\}, \mathscr{S}_1, \ldots, \mathscr{S}_{L-1}$ that are nested, in the sense that each $S \in \mathscr{S}_\ell$ can be written as an union of sets in $\mathscr{S}_{\ell+1}$, $0 \le \ell \le L-2$. Also, by recursion, we ensure the following property:

(a) for all $\ell \in \{1, \ldots, L-1\}$,
$$\#\mathscr{S}_\ell / \#\mathscr{S}_{\ell-1} \le 2(2em\ell N_\ell/\bar{s}_\ell)^{\bar{s}_\ell}.$$

(b) for all $\ell \in \{1, \ldots, L-1\}$, all $S \in \mathscr{S}_\ell$, all $j \in \{1, \ldots, m\}$, the output of any neuron of the ℓ-th layer of the network taking x_j as input is a multivariate polynomial function of $w \in S$, of degree at most ℓ.

Let us start with $\ell = 1$, which already contains the key idea. Consider the mN_1 functions $\Psi_{h,j}$, $1 \le h \le N_1$, $1 \le j \le m$, as function of $w \in \mathbb{R}^{\bar{s}_1}$, where $\Psi_{h,j}$ corresponds to the output of the h-th neuron of the hidden Layer when the network takes as input x_j. They have the form $(p_{h,j})_+$, where (recall that the input dimension of the network is 1),

$$p_{h,j} : w = (a_h, b_h)_{1 \le h \le N_1} \in \mathbb{R}^{\bar{s}_1} \mapsto a_h x_j + b_h \in \mathbb{R}.$$

Now it is clear that *all* the functions $(p_{h,j})_+$ are *simultaneously* multivariate polynomial functions of w (of degree at most 1) provided that w is restricted to a subset of $\mathbb{R}^{\bar{s}_1}$ that makes the mN_1-dimensional vector sign $(p_{j,h}(w), 1 \le h \le N_1, 1 \le j \le m)$ unchanged. Hence, we can build a partition $\mathscr{S}_1$ of $\mathbb{R}^{\bar{s}_1}$ with Property (b) by looking at the portions of the space $\mathbb{R}^{\bar{s}_1}$ that are shaped with the mN_1 constraints $\text{sign}(a_h x_j + b_h) = \pm 1$, $1 \le h \le N_1, 1 \le j \le m$. More formally,

$$\mathscr{S}_1 = \left\{ \bigcap_{1 \le h \le N_1} \bigcap_{1 \le j \le m} (\text{sign} \circ p_{h,j})^{-1}(\varepsilon_{j,h}), \varepsilon \in \{-1, +1\}^{mN_1} \right\}.$$

[1] Here we allows for empty elements in the partition.

To check (a), we should now evaluate the cardinal of this partition. In fact, it is

$$\#\{(\text{sign} \circ p_{j,h}(w), 1 \le h \le N_1, 1 \le j \le m), w \in \mathbb{R}^{\bar{s}_1}\}$$

which is bounded by $2(2emN_1 \times 1/\bar{s}_1)^{\bar{s}_1}$ by using Lemma 3.3 (there are mN_1 polynomials, with degree 1, that are functions of $\bar{s}_1$ variables). This yields (a) and (b) for $\ell = 1$. (Note that, strictly, $\mathscr{S}_1$ is a partition of $\mathbb{R}^{\bar{s}_1}$, but it can be straightforwardly extended to a partition of $\mathbb{R}^s$ by multiplying each element of the partition by the set $\mathbb{R}^{s-\bar{s}_1}$).

Now assume that $\mathscr{S}_1, \ldots, \mathscr{S}_{\ell-1}$ are built with (a) and (b), and let us build the new partition $\mathscr{S}_\ell$ for layer ℓ. Fix S a member of the partition $\mathscr{S}_{\ell-1}$. We consider the mN_ℓ functions $\Psi_{h,j}$, $1 \le h \le N_\ell$, $1 \le j \le m$, as function of $w \in \mathbb{R}^{\bar{s}_\ell}$, where $\Psi_{h,j}$ corresponds to the output of the h-th neuron of the ℓ-th layer when the network takes as input x_j. They have the form $(p_{h,j})_+$, where

$$p_{h,j} : w \in S \mapsto \sum_{k=1}^{N_{\ell-1}} a_h^k s_{k,\ell}(w) + b_h \in \mathbb{R}.$$

where $s_{k,\ell}(w)$ is the output of the k-th neuron of the layer $\ell-1$ (for the input x_j of the network), which is a multivariate polynomial of degree $\ell-1$ in $w \in S$. Hence, $p_{h,j}$ is a multivariate polynomial of degree ℓ in $w \in S$. Applying the same reasoning as above (with Lemma 3.3), we can further partition the set S into at most $2(2emN_\ell\ell/\bar{s}_\ell)^{\bar{s}_\ell}$ elements. This gives a new partition satisfying (a) and (b).

Finally, we consider the partition $\mathscr{P} = \mathscr{S}_{L-1}$ of $\mathbb{R}^s$. By (b), the output of each neuron of the $(L-1)$-th layer (for the input x_j of the network) is a polynomial of degree at most $L-1$ in $w \in S$, for each $S \in \mathscr{P}$. Hence, the output of the network, for each input x_j, is a polynomial of degree at most L in $w \in S$, for each $S \in \mathscr{P}$. By (a), the cardinal of the partition is bounded by

$$\prod_{\ell=1}^{L-1} 2(2emN_\ell\ell/\bar{s}_\ell)^{\bar{s}_\ell}.$$

This yields the claim.

3.2 Entropy of a class of neural networks

This section is based on the paper [SH20] (see also references therein).

3.2.1 Covering number and entropy

Let us start by recalling the general definition of the covering number, see, e.g., [SSBD14, MRT18].

Definition 3.6. *Let E be a vector space, endowed with a norm $\|\cdot\|$, and $A \subset E$. For any $\delta \in (0,1)$, the covering number $\mathscr{N}(\delta, A, \|\cdot\|)$ of A is the minimum number of $\|\cdot\|$-balls with radius δ that covers A. More formally,*

$$\mathscr{N}(\delta, A, \|\cdot\|) = \min\left\{k \ge 1 : \exists e_1, \ldots, e_k \in E,\ A \subset \cup_{i=1}^k B_{\|\cdot\|}(e_i, \delta)\right\}, \tag{3.7}$$

where $B_{\|\cdot\|}(e_i, \delta) = \{x \in E : \|x - e_i\| \le \delta\}$. The entropy of A is then defined by $\log(\mathscr{N}(\delta, A, \|\cdot\|))$.

Let us provide some examples:

- For $E = \mathbb{R}$ endowed with the absolute value $\|\cdot\| = |\cdot|$, any subset $A \subset [-c, c]$ is such that

  $$\mathscr{N}(\delta, A, \|\cdot\|) \le 2c/\delta.$$

  Indeed, letting $k = \lfloor 2c/\delta \rfloor$, and letting $(x_1, \ldots, x_k) = (-c+\delta, -c+2\delta, \ldots, -c+k\delta)$, we have that any point of $[-c, c]$ is at distance less than δ of one of the x_i's (note that $-c + k\delta > c - \delta$ by definition of k).

- For $\mathbb{E} = \mathbb{R}^d$ with the infinite norm $\|\|_\infty$, any subset $A \subset [-1,1]^d$ is such that

$$\mathcal{N}(\delta, A, \|\cdot\|) \leq (2/\delta)^d.$$

This is the same idea as above, transposed in dimension d: we produce a grid of $[-1,1]^d$ with mesh size δ. Letting $k = \lfloor 2/\delta \rfloor$, this can be done by considering the grid $(x_1, \ldots, x_k) = (-1+\delta, -1+2\delta, \ldots, -1+k\delta)$ on all dimensions. The number of points in this grid is k^d.

- Still for $(\mathbb{R}^d, \|\|_\infty)$, a subset that will be useful in the sequel is the set

$$A = \{x \in \mathbb{R}^d : \|x\|_0 \leq s, \|x\|_\infty \leq 1\}, \tag{3.8}$$

containing s-sparse vectors[2], for some sparsity $1 \leq s \leq d$. In that case, we have

$$\mathcal{N}(\delta, A, \|\cdot\|) \leq (2d/\delta)^{s+1}. \tag{3.9}$$

Indeed, we have $A \subset \cup_{S \subset \{1,\ldots,d\}, |S| \leq s} A_S$, where $A_S = \{x \in \mathbb{R}^d : \forall i \notin S, x_i = 0, \forall i \in S, |x_i| \leq 1\}$. By the above case, A_S can be covered with $\leq (2/\delta)^{|S|}$ $\|\cdot\|_\infty$-balls of radius δ (extend the centers of the ball in $\mathbb{R}^S$ to $\mathbb{R}^d$ by adding the appropriate number of 0 coordinates). Hence, we can cover A with a number of such balls smaller than

$$\sum_{r=0}^{s} \binom{d}{r} (2/\delta)^r \leq \sum_{r=0}^{s} d^r (2/\delta)^r \leq (2d/\delta)^{s+1}/(2d/\delta - 1) \leq (2d/\delta)^{s+1},$$

because $2d/\delta \geq 2$.

Note that the dependence in δ of the bound (3.9) is much smaller than bound (3.8) when $s \ll d$.

In the sequel, we consider E being the vector space of bounded functions from $[0,1]^d$ in $\mathbb{R}$, endowed with the infinite norm. For some function set $\mathscr{F}$, the covering number $\mathcal{N}(\delta, \mathscr{F}, \|\cdot\|_\infty)$ is well known to be a key tool to evaluate the stochastic error of the empirical risk minimizer (ERM) in regression, as recalled by the next result.

Theorem 3.7 (Lemma 4 in [SH20])**.** *Let us consider the non-parametric regression model where $(X_i)_{1 \leq i \leq n}$ are i.i.d. copy of X, some random variable valued in $[0,1]^d$, and the responses are $Y_i = f_0(X_i) + \varepsilon_i$, $1 \leq i \leq n$, where $(\varepsilon_i)_{1 \leq i \leq n}$ are i.i.d. $\mathcal{N}(0,1)$ (independent of the X_i's). Let $\mathscr{F}$ be a class of functions from $[0,1]^d$ to $[-M, M]$ (for some $M > 0$) and $\hat{f}$ be the empirical risk minimizer over this class, that is,*

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

Then we have for any $f_0 : [0,1]^d \to [-M, M]$, for all $\delta, \epsilon > 0$,

$$\mathbb{E}\left[(\hat{f}(X) - f_0(X))^2\right] \leq (1+\epsilon)\left[\inf_{f \in \mathscr{F}} \mathbb{E}\left[(f(X) - f_0(X))^2\right] + M^2 \frac{18 \log(\mathcal{N}) + 72}{n\epsilon} + 32\delta M\right],$$

for which $\mathcal{N} = \mathcal{N}(\delta, \mathscr{F}, \|\cdot\|_\infty) \geq 3$.

The proof will be investigated in Chapter 4.

[2] Let us recall that a vector is s-sparse if it has no more than s non-zero coordinates.

3.2.2 Result

The aim of this section is to provide an upper bound on the covering number of a function set defined by realizations of NN. We consider ρ being the ReLU activation (although we will only use that ρ is a 1–Lip function with $\rho(0) = 0$). Since ρ is fixed throughout, we write $R(W)$ for $R(\Phi)$.

Let us first consider the set of NN realizations from $\mathbb{R}^d$ to $\mathbb{R}$ with weights bounded by 1, that is, with the notation of Chapter 1,

$$\mathscr{F}(L,N) = \big\{f = R(\Phi), \text{ for some } (A_\ell)_{1\le\ell\le L}, (b_\ell)_{1\le\ell\le L}, \quad N_0 = d, \quad N_L = 1, \quad \max_{1\le\ell\le L} (\|A_\ell\|_\infty \vee |b_\ell|_\infty) \le 1\big\}.$$

Next, for $s > 0$ being a sparsity parameter, we let

$$\mathscr{F}(L,N,s) = \Big\{f \in \mathscr{F}(L,N), \quad \sum_{\ell=1}^{L} (\|A_\ell\|_0 + |b_\ell|_0) \le s\Big\}.$$

It is important to note that the support of the network *is not fixed here*, only the number of non-zero weights is assumed to be below s. Hence, compared to the NN considered in Section 3.1, the topology of the considered NN is much less constrained, which is more suitable for optimization. As a counterpart, the weights are assumed to be bounded here. However, taking bounded weights is a classical recommendation when using NN in practice, so it is not a big assumption here.

Finally, we denote

$$V = V(N) := \prod_{\ell=0}^{L} (N_\ell + 1). \tag{3.10}$$

Then the following result holds.

Theorem 3.8. *For V as in (3.10) and any $\delta \in (0,1)$, we have*

$$\log \mathscr{N}(\delta, \mathscr{F}(L,N,s), \|\cdot\|_\infty) \le (s+1)\log\left(\frac{2LV^2}{\delta}\right).$$

This bound can in turn be used in results like Theorem 3.7 to upper bound the stochastic error term. Obviously, this has to balanced with the bias error term in order to provide a bound on the generalization risk. This will be investigated in the next chapter.

3.2.3 Proof of Theorem 3.8

Let us denote by T the total number of parameters of the NN, that is,

$$T = \sum_{\ell=1}^{L} (N_\ell N_{\ell-1} + N_\ell).$$

Hence, T is the number of *possibly active* parameters of NN in $\mathscr{F}(L,N,s)$, while s is the number of *actually active* parameters. In the sequel, we can thus describe the set $\mathscr{F}(L,N,s)$ as the functions h_w, with $w \in \mathbb{R}^T$ which is s-sparse and $\|w\|_\infty \le 1$.

Applying the covering number bound (3.9), for any $\epsilon \in (0,1)$, there exist $w_1, \dots, w_k$ with $k \le (2T/\epsilon)^{s+1}$ such that

$$\mathscr{F}(L,N,s) \subset \bigcup_{i=1}^{k} \{h_w : \|w - w_i\|_\infty \le \epsilon\}.$$

Now, applying Lemma 3.9, we have that $\|w - w_i\|_\infty \le \epsilon$ entails $\|h_w - h_{w_i}\|_\infty \le \epsilon LV$. Hence, we deduce

$$\mathscr{F}(L,N,s) \subset \bigcup_{i=1}^{k} \{h : \|h - h_{w_i}\|_\infty \le \epsilon LV\}.$$

This means that for any $\delta \in (0,1)$, we have

$$\mathcal{N}(\delta, \mathscr{F}(L,N,s), \|\cdot\|_\infty) \le (2T/\epsilon)^{s+1},$$

for $\epsilon = \delta/(LV)$. Since

$$T \le \sum_{\ell=1}^{L} (N_{\ell-1}+1)(N_\ell+1) \le \sum_{\ell=1}^{L} 2^{-L+1} V \le (2L2^{-L})V \le V,$$

because for all $x \ge 2$, $2x2^{-x} \le 1$. We finally obtain

$$\mathcal{N}(\delta, \mathscr{F}(L,N,s), \|\cdot\|_\infty) \le (2LV^2/\delta)^{s+1},$$

which is the desired bound.

3.2.4 A useful Lemma

The next lemma quantifies how much small errors in network parameters propagate into a global error for the network realisation.

Lemma 3.9. *Suppose $f = R(W)$ and $f^* = R(W^*)$ belong to $\mathscr{F}(L,N)$ with $W = (A_k, b_k)_{k=1,\dots,L}$ and $W^* = (A_k^*, b_k^*)_{k=1,\dots,L}$. Suppose that individual entries of A_k's and b_k's are at most $\varepsilon > 0$ away from the corresponding entries of A_k^* and b_k^*. Then for V as in (3.10),*

$$\|f - f^*\|_\infty \le \varepsilon LV.$$

Proof. Recall $f = T_L \circ \rho \circ \cdots \circ \rho \circ T_1$ with $T_k(x) = A_k x + b_k$ and define, for $k = 1, \dots, L$,

$$B_k f = \rho \circ T_k \circ \cdots \circ \rho \circ T_1,$$
$$E_k f = T_L \circ \rho \circ \cdots \circ T_{k+1} \circ \rho,$$

and set $E_L f = B_0 f = \mathrm{Id}$. We first prove two basic facts about $B_k f, E_k f$.

Fact 1. If $f \in \mathscr{F}(L,N)$, then $|(B_k f)(x)|_\infty \le \prod_{l=1}^{k} (N_{l-1}+1)$ for $x \in [0,1]^d$.

Let us check first that $|(\rho \circ T_i)(y)|_\infty \le N_{i-1}|y|_\infty + 1$ for any integer i. Indeed, $|\rho(y)|_\infty \le |y|_\infty$ and $|T_k y|_\infty \le |A_k y|_\infty + |b_k|_\infty \le N_{k-1}|y|_\infty + 1$, using $\|A_k\|_\infty \le 1, |b_k|_\infty \le 1$. In particular, if $|y|_\infty \ge 1$ we have $|(\rho \circ T_i)(y)|_\infty \le (N_{i-1}+1)|y|_\infty$ for any i.

The result follows by recursion: for $i = 1$ we get $|(\rho \circ T_1)(x)|_\infty \le N_0|x|_\infty + 1 \le N_0 + 1$. Since $N_0 + 1 \ge 1$ it suffices feeds this bound into the previous inequality in terms of y.

Fact 2. The map $x \to (E_k f)(x)$ is Λ_k–Lipschitz, with $\Lambda_k \le \prod_{l=k+1}^{L} N_{l-1}$.

The composition of an L_1–Lip by an L_2–Lip function is an $L_1 L_2$–Lip function. By definition ρ is 1–Lip, while T_i is N_{i-1}–Lip for any i, from which the fact follows.

Now let us write the difference $f - f^*$ as the telescopic sum

$$f(x) - f^*(x) = \sum_{k=1}^{L} \Big[(E_k f) \circ T_k \circ (B_{k-1} f^*)(x) - (E_k f) \circ T_k^* \circ (B_{k-1} f^*)(x) \Big].$$

Combining the triangle inequality with Fact 2 above,

$$\begin{aligned}|f(x)-f^*(x)| &\leq \sum_{k=1}^{L} \Lambda_k \left|(T_k - T_k^*)\circ(B_{k-1}f^*)(x)\right|_\infty \\ &\leq \sum_{k=1}^{L} \Lambda_k \left[\|A_k - A_k^*\|_\infty |(B_{k-1}f^*)(x)|_1 + |b_k - b_k^*|_\infty\right] \\ &\leq \sum_{k=1}^{L} \Lambda_k \left[\varepsilon N_{k-1}|(B_{k-1}f^*)(x)|_\infty + \varepsilon\right].\end{aligned}$$

The term under brackets in the last display is at most, using Fact 1,

$$\varepsilon N_{k-1}\prod_{l=1}^{k-1}(N_{l-1}+1)+1 \leq \varepsilon \prod_{l=1}^{k}(N_{l-1}+1).$$

One deduces the announced result

$$|f(x)-f^*(x)| \leq \varepsilon \sum_{k=1}^{L}\prod_{l=1}^{L}(N_{l-1}+1) \leq \varepsilon LV.$$

□

Bibliography

[ABB+99] Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

[BHLM19] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

[BSH21] Thijs Bos and Johannes Schmidt-Hieber. Convergence rates of deep relu networks for multiclass classification. *arXiv preprint arXiv:2108.00969*, 2021.

[Ger21] Sébastien Gerchinovitz. Fondements théoriques de l'apprentissage profond. chapitre 6. https://www.math.univ-toulouse.fr/~fmalgouy/enseignement/mva.html, 2021.

[Gir15] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[SH20] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.