Chapter 4

Regression

Contents

4.1 The model

Consider observing i.i.d. pairs $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ with

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad 1 \le i \le n, \tag{4.1}$$

where X_i are $[0,1]^d$–valued random variables (also called *design points*) and ε_i are independent standard normal $\mathcal{N}(0,1)$ variables, and independent of the X_i's, and $f_0 : [0,1]^d \to \mathbb{R}$ an unknown function.

Typical statistical goals in this setting are

- estimating the unknown regression function f_0 from the observations
- finding estimates that behave (near–)"optimally" with respect to some criterion (e.g. minimax) over natural classes of parameters.

Let $\hat{f}(\cdot) = \hat{f}_n(Z_1, \ldots, Z_n)(\cdot)$ be an estimator of f.

The *prediction* risk in the setting of model (4.1) is defined as follows. Let T be a 'synthetic' data point, that is a variable independent of the X_i's and generated from the distribution of X_1. Let

$$R(\hat{f}, f_0) = E\left[\left(\hat{f}(T) - f_0(T)\right)^2\right] = E\left[\left(\hat{f}(Z_1, \ldots, Z_n)(T) - f_0(T)\right)^2\right]. \tag{4.2}$$

The interpretation is as follows: from the observations $Z_1, \dots, Z_n$, one builds an estimator $\hat{f}(Z_1, \dots, Z_n)(\cdot)$, and then evaluate its performance 'on average' for the prediction of the value of f_0 at a new point T. More explicitly, since $Z_1, \dots, Z_n$ are iid with distribution P_{Z_1} the law of Z_1, and T has same law as X_1,

$$R(\hat{f}, f_0) = \int \cdots \int \left(\hat{f}(z_1, \dots, z_n)(t) - f_0(t)\right)^2 dP_{Z_1}^{\otimes n}(z_1, \dots, z_n) dP_{X_1}(t).$$

The results of the chapter also hold for the following 'empirical' risk, where the L^2–norm is replaced by the empirical L^2–norm:

$$\hat{R}(\hat{f}, f_0) = E[\|\hat{f} - f_0\|_n^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f_0(X_i))^2\right], \tag{4.3}$$

with $\hat{f}(X_i) = \hat{f}(X_i)$ the value of $\hat{f}$ at point X_i, that is, if one uses a fully explicit notation, $\hat{f}(Z_1, \dots, Z_n)(X_i)$ (note that unlike the notation could perhaps suggest, the quantity $\hat{R}(\hat{f}, f_0)$ is non–random). Results for this risk are obtained as byproduct of the proofs below.

4.2 Near-optimal estimation rates with ReLU–DNNs

Class of smooth functions. Let us first recall that a function g is β–Hölder on $[0,1]$, with $\beta \in (0,1]$, if

$$\sup_{x,y\in[0,1], x\neq y} \frac{|g(x) - g(y)|}{|x - y|^\beta} < \infty.$$

Next to get appropriate 'balls' of Hölder functions, one bounds the ratio in the previous display by a finite constant, say K, as well as bounds the supremum norm of g, again by K, or the sum of the two quantities by K (without this second constraint one could add an arbitrary constant to g, while in practice it seems reasonable to assume that g is bounded). We summarise these constraints by now giving the general definition of a Hölder ball in dimension d that we consider in the sequel.

A Hölder ball of functions on $[0,1]^d$ is defined as, for $\beta > 0$ and $K > 0$,

$$\mathscr{C}^\beta([0,1]^d, K) = \left\{ f : [0,1]^d \to \mathbb{R} : \sum_{\boldsymbol{\alpha}:\, \sum_{i=1}^d \alpha_i < \beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:\, \|\boldsymbol{\alpha}\|_1 = \lfloor\beta\rfloor} \sup_{x,y\in[0,1]^d,\, x\neq y} \frac{|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}} f(y)|}{\|x - y\|_\infty^{\beta - \lfloor\beta\rfloor}} \leq K \right\},$$

where $\lfloor\beta\rfloor$ *here* denotes the largest integer strictly smaller than β (so that $\lfloor 1\rfloor = 0$) and where a multi-index notation is used for the partial derivative $\partial^{\boldsymbol{\alpha}} f = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$, with $(\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$.

Target rate. The notion of *optimal estimation rate* is defined with respect to the minimax criterion and the prediction risk. The minimax risk over a class $\mathscr{C}$ of functions – for instance $\mathscr{C} = \mathscr{C}_d^\beta([0,1]^d, K)$ – for estimation of f in model (4.1) is defined as

$$R_M = \inf_{\hat{f}} \sup_{f_0 \in \mathscr{C}} E[(\hat{f}(T) - f_0)^2],$$

where the infimum is over all possible estimators of f_0 in model (4.1).

It is a standard result that when $\mathscr{C} = \mathscr{C}_d^\beta([0,1]^d, K)$, then as $n \to \infty$,

$$R_M \asymp n^{-\frac{2\beta}{2\beta + d}}.$$

We refer to a course on nonparametric estimation for more details, where this rate is typically established for regression and related models such as density estimation. There are several popular classes

of estimators that achieve this global rate (sometimes up to logarithmic factors), such as wavelet estimators, kernel estimators, Bayesian posterior distributions etc.

One may note that the rate becomes very slow if d is large (unless perhaps for very high smoothness levels; for example if one imagines that d is allowed to grow as $(\log n)$, then the rate becomes constant). This is sometimes referred to as the *curse of dimensionality*. This is not a problem that comes from the use of a particular estimation method: *any* method will have a slow rate in dimension d large, when applied with certain 'un-favourable' true regression functions. This problem rather comes from the fact that the Hölder class becomes very massive as the dimension d increases. We will come back to the *curse of dimensionality* later in the chapter, where it will be seen that if the true function f_0, although defined on the whole of $\mathbb{R}^d$, has an actual small 'effective' dimension, the deep network estimator can naturally adapt to this smaller dimension.

4.2.1 Deep ReLU estimator

A natural idea to find an estimator of the regression function f in model (4.1) is to do 'maximum likelihood'. Since the errors ε_i are independent normal, this is the same as 'doing least squares'. So, given a class $\mathscr{F}$ of functions to be chosen below and observations $(X_i, Y_i)_i$ from model (4.1), let us set

$$\hat{f}^{ERM} = \hat{f}^{ERM}(\mathscr{F}) = \operatorname*{argmin}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2. \tag{4.4}$$

That is, one wants to find the 'best fit' to the data over the class $\mathscr{F}$ in terms of least squares. This estimator is called the Empirical Risk Minimizer over the class $\mathscr{F}$ (in short ERM).

Let us immediately note that in practice one often does not know an exact minimizer as in (4.4) but one can only compute some approximation of it $\tilde{f}$. How 'far' this practical estimator is from the ERM can be measured through

$$\Delta_n(\tilde{f}, f_0) = E_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{f}(X_i))^2 - \inf_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2\right].$$

Within this Chapter we consider the properties of the ERM itself, i.e. we assume that we have access to the ERM (or of a good enough approximation thereof, so that the term Δ_n is negligible compared to the main convergence rate term). More discussion about the term Δ_n will be given within the Interpolation Chapter of this course. Below we set $a\wedge b = \min(a,b)$ and $a\vee b = \max(a,b)$.

Definition 4.1 (Classes of DNN networks)**.** *Consider a NN Φ with input dimension d, output dimension 1, depth L, width vector $N = (N_l)_{1\le l\le L}$ and activation ρ*

$$\Phi = ((A_1, b_1), \dots, (A_L, b_L))$$

Define a 'ball' of network realisations, with parameters bounded by 1, as

$$\mathscr{F}(L, N) = \big\{f = R(\Phi),\ \text{for some } (A_j)_{1\le j\le L}, (b_j)_{1\le j\le L},\quad \max_{1\le j\le L}\big(\|A_j\|_\infty \vee |b_j|_\infty\big) \le 1\big\}.$$

We also set, for $s > 0$ a sparsity parameter,

$$\mathscr{F}(L, N, s, F) = \Big\{f \in \mathscr{F}(L, N),\quad \sum_{j=1}^{L}(\|A_j\|_0 + |b_j|_0) \le s,\quad \|f\|_\infty \le F\Big\}$$

Definition 4.2 (Deep ReLU estimator)**.** *For $\mathscr{F} = \mathscr{F}(L, N, s, F)$ as in Definition 4.1 for some $L, N, s, F > 0$, let us set*

$$\hat{f} = \hat{f}^{ReLU} = \operatorname*{argmin}_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2. \tag{4.5}$$

That is, $\hat{f}$ is the ERM with optimisation over the set $\mathscr{F}(L, N, s, F)$ of s–sparse ReLU neural networks.

We assume here that the network parameters are bounded in absolute value by 1. Another positive constant could be used. The rationale behind this choice is that in practice most of the time network parameters are initialized using bounded parameters. It is also typically observed that even after training the parameters of the network remain quite close in range to initial parameters. From the theoretical point of view, some approximation results are quite easily reachable if network parameters are allowed to be very large. But in order to be closer to practical applications where parameters typically remain bounded, we restrict ourselves to that case, and we will see below that this does not prevent us to obtain good inference properties.

4.2.2 Global convergence rate for Hölder functions

The following result shows that deep ReLU neural network can achieve the optimal rate $\varepsilon_n^2 = n^{-2\beta/(2\beta+d)}$ in terms of prediction loss in regression, up to logarithmic factors, provided the parameters of the class of networks $\mathscr{F}$ over which the ERM is computed (4.5) are well–chosen.

Theorem 4.3. *Suppose the true unknown $f_0 \in \mathscr{C}^\beta([0,1]^d, K)$ for an arbitrary $\beta > 0$ and $K > 0$. 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 $N = (N_l)_{1\le l\le L}$, sparsity s and uniform bound F. Suppose $F \ge K \vee 1$ and a choice of parameters as follows*

$$\log n \lesssim L \le n^{\frac{d}{2\beta+d}}, \qquad n^{\frac{d}{2\beta+d}} \le \min_{1\le l\le L} N_l \le \max_{1\le l\le L} N_l \le n^2, \qquad s \asymp (\log n) n^{\frac{d}{2\beta+d}}.$$

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

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

Let us now discuss possible choices of L, N, s in more details. From Theorem 4.3 it appears that the network should not be too deep, as the depth is in factor of the target rate. On the other hand, it will be clear from the proof that a depth of at least logarithmic order allows for approximation of smooth functions by the ReLU network realisation at a polynomial rate in n^{-1} (see e.g. Lemmas 4.8 and 4.9). So it seems natural to choose $L \asymp \log n$. The widths of the network, on the other hand, can be chosen to be polynomial in n, for instance one can take $N_l \asymp n$ (the upper–bound n^2 can be replaced by any power of n). Finally, the perhaps most important parameter in terms of sensibility of the rate is the sparsity s: it should be finely tuned in terms of the true regularity β of f_0 as $n^{1/(2\beta+d)}$ (up to a log factor). This parameter plays a similar role as e.g. the 'bandwidth' parameter of a kernel estimator: note that in fact the choice is similar to the one of an optimal bandwidth for kernel regression.

Corollary 4.4. *Under the same setting as Theorem 4.3, if one optimises over networks with depth $L \asymp \log n$, all widths $N_l \asymp n$ and sparsity $s \asymp (\log n) n^{\frac{1}{2\beta+1}}$, then the ERM $\hat{f}$ in (4.5) verifies*

$$\sup_{f_0 \in \mathscr{C}^\beta([0,1]^d, K)} R(\hat{f}, f_0) \lesssim (\log n)^3 n^{-\frac{2\beta}{2\beta+d}}.$$

Note that statistically, the previous results are still 'non-adaptive', that is, in order to use the previous parameters one needs to know the regularity of f_0, which is rarely the case in practice. One may envision coupling the ERM with one of the adaptation methods from the standard nonparametrics toolbox, for instance a penalisation method. Alternatively, one could use a Bayesian method (instead of the ERM), for which adaptation (in L^2 losses) is often relatively straightforward to obtain.

Let us now underline remarkable consequences of these results. We have seen in Chapter 1 that deep ReLU networks realisations are Lipschitz functions, and it has been seen in Chapter 2 that deep ReLU networks have good approximation properties over the class of Lipschitz functions.

- *Smoothness.* We see that deep ReLU empirical risk minimizers can attain (up to a poly-log factor) the optimal minimax rate for Hölder classes for *any* smoothness parameter $\beta > 0$, not only regularities between 0 and 1, as would be the case for regular histogram estimators, or between 0 and 2 as would be the case for classical piecewise affine 'local polynomial' estimators. The flexibility of deep ReLU networks comes from the compositional structure: with a small number of compositions, it enables to estimate quickly smooth functions, although the overall resulting realisation is still piecewise constant.
- *Adaptation to hidden structures.* The convergence rate obtained in Theorem 4.3 is only an upper-bound, which can be quite pessimistic; it turns out that in many cases, the actual rate attained by the deep ReLU estimator is much faster, at least if the 'intrinsic' dimensionality of the regression function is smaller than the ambient space dimensionality d. This will be studied in Section 4.4.
- *Weights all between* 0 *and* 1. The optimisation set $\mathscr{F}$ for the network parameters assumes bounded weights, between 0 and 1: it is interesting to see that a near-optimal estimation rate can be achieved without taking weights possibly growing with the sample size.

4.2.3 Key ideas

The proof of Theorem 4.3 is based on two important sub-results. The first is of deterministic nature and deals with approximation of smooth functions through ReLU DNNs.

Theorem 4.5 (Approximation of smooth functions by DNNs)**.** *Let $f \in \mathscr{C}_d^\beta([0,1]^d, K)$ be a function of regularity $\beta > 0$. Let $m, \mathcal{N} \geq 1$ be two integers. There exists a network, with $\Lambda := 6(d + \lceil\beta\rceil)\mathcal{N}$,*

$$\tilde{f} \in \mathscr{F}(L, (d, \Lambda, \dots, \Lambda, 1), s, \infty)$$

with depth and sparsity verifying, for c_0, C_0 depending on d, β only,

$$L = C_0 m, \qquad s \leq c_0 m \mathcal{N},$$

such that, for $c_1, c_2, \mathcal{N}_0$ depending on d, β, K only, and all $\mathcal{N} \geq \mathcal{N}_0$,

$$\|\tilde{f} - f\|_\infty \leq c_1 \frac{\mathcal{N}}{4^m} + c_2 \mathcal{N}^{-\frac{\beta}{d}}.$$

The second result is a general oracle inequality that is valid for any empirical risk minimizer. We apply it below to the DNN estimator $\hat{f}$ in (4.5).

Theorem 4.6 (Lemma 4 in [SH20b])**.** *Let $\mathscr{F}$ be a class of functions from $[0,1]^d$ to $[-F, F]$ (for some $F \geq 1$) and $\hat{f}$ be the empirical risk minimizer over this class in the regression model (4.1), 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 $f_0 : [0,1]^d \to [-F, F]$, for all $\delta, \varepsilon > 0$,

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

for which $\mathcal{N}_n(\delta) := \mathcal{N}(\delta, \mathscr{F}, \|\cdot\|_\infty)$ and assuming $3 \leq \mathcal{N}_n(\delta) \leq e^n$.

The upper-bound for the prediction risk in Theorem 4.6 can be interpreted as displaying a bias–variance trade–off. The first term corresponds to *bias*: it measures the best possible approximation from the given candidate class $\mathscr{F}$. The term displaying the entropy can be interpreted as complexity or 'variance', measuring the complexity of the class $\mathscr{F}$ over which the ERM is taken.

Both results will turn useful again when analysing different assumptions on the function class the true f_0 belongs to in Section 4.4.

4.3 Proof of the global rate theorem

Let us first see how Theorem 4.3 follows by combining Theorems 4.5 and 4.6. The later oracle inequality also needs a quantitative bound on the entropy, which is provided in Lemma 4.13 below.

One applies Theorem 4.6 with $\mathscr{F} = \mathscr{F}(L, N, s, F)$ the class of DNN realisations from Theorem 4.3 with parameters as specified in the statement. One can bound

$$\inf_{f\in\mathscr{F}} E\left[(f(T) - f_0(T))^2\right] \leq \inf_{f\in\mathscr{F}} \|f - f_0\|_\infty^2.$$

Setting $\delta = 1/n, \varepsilon = 1$ and using Theorem 4.6,

$$E\left[(\hat{f}(T) - f_0(T))^2\right] \lesssim \inf_{f\in\mathscr{F}} \|f - f_0\|_\infty^2 + F^2 \frac{18\log \mathscr{N}_n(1/n) + 72}{n} + 32\frac{F}{n}.$$

Using Lemma 4.13 and recalling $V \leq (2\max N_l)^L \leq (2n^2)^L$ via (4.15) and the conditions on N_ls,

$$\log \mathrm{N}(\delta, \mathscr{F}(L, N, s), \|\cdot\|_\infty) \leq (s+1)\log\left(\frac{2LV^2}{\delta}\right) \leq 2s[\log(2Ln) + 2L\log(2n^2)],$$

which is bounded by $CsL\log n$. Now to control the term with the infimum above, we apply Theorem 4.5 with $\mathscr{N} \asymp n^{\frac{d}{2\beta+d}}$, $m = \lfloor \log_2 n \rfloor$. There exists a network in $\mathscr{F}(L', (d, \Lambda, \ldots, \Lambda, 1), s, \infty)$, with $L' = m = \lfloor \log_2 n \rfloor$ and $s \lesssim m\mathscr{N} \asymp (\log n) n^{\frac{d}{2\beta+d}}$ such that its realisation $\tilde{f}$ verifies

$$\|\tilde{f} - f_0\|_\infty \lesssim \frac{\mathscr{N}}{4^m} + \mathscr{N}^{-\frac{\beta}{d}} \lesssim n^{-\frac{\beta}{2\beta+d}}.$$

Note that by embedding between spaces (a network of given depth can always be embedded into one with larger depth up to slightly updating the overall sparsity), there is a network with realisation $\check{f}$ and parameters as in the statement of the theorem (in particular, depth L) such that the previous display holds with $\check{f}$ instead of $\tilde{f}$. Putting the previous bounds together leads to

$$E\left[(\hat{f}(T) - f_0(T))^2\right] \lesssim n^{-\frac{2\beta}{2\beta+d}} + \frac{sL\log n}{n} + \frac{1}{n} \lesssim L(\log n)^2 n^{-\frac{2\beta}{2\beta+d}},$$

which concludes the proof of Theorem 4.3.

4.3.1 Ingredient 1: smooth approximation with deep ReLU networks

Let us now prove Theorem 4.5. The general idea is as follows: there are two main steps. The first is not specific to DNNs and is that any β–Hölder function can be well–approximated locally, using Taylor expansions, by a polynomial of order $\lfloor\beta\rfloor$: one can approximate f_0 by a piecewise polynomial function, with a quality of approximation that depends on β. The second idea, where the choice of activation function σ comes in, is that it is possible to approximate quickly, in one dimension, the