Statistical Theory for Deep Learning

Lecture notes

Prof. Dr. Michael Kohler

Fachbereich Mathematik

Technische Universität Darmstadt

Sommersemester 2024

Contents

1 Introduction **3**

1.1 Feedforward neural networks 3

1.2 Convolutional neural networks 5

1.3 Transformer networks 7

1.4 Learning of deep (neural) networks 10

2 Neural network approximation **14**

2.1 Introduction 14

2.2 Approximation power of deep neural networks 17

3 Neural network generalization **28**

3.1 Motivation 28

3.2 Uniform exponential inequalities 29

3.3 Covering numbers and VC dimension 43

3.4 VC dimension of sets of deep neural networks 54

4 Least squares neural network regression estimates **59**

4.1 A general result 59

4.2 Rate of convergence of least squares neural network estimates 61

4.3 Lower bounds on the rate of convergence 63

4.4 Deep learning as a remedy against the curse of dimensionality 71

1 Introduction

Deep Learning are methods, which fit *deep (neural) networks* to data. These methods have achieved tremendous success in applications in the last ten years in the areas

1. mastering of games (AlphaGo from Google DeepMind won in 2015 against European champion in Go),
2. image classification,
3. text classification,
4. machine translation,
5. simulation of human conversation (ChatGPT).

The reason for this success is twofold:

1. availability of huge data sets,
2. massive increasement of computing power.

The nowadays most popular deep (neural) networks are

1. feedforward neural networks,
2. convolutional neural networks,
3. transformer networks,

which will be introduced next.

1.1 Feedforward neural networks

Feedforward neural networks are biological motivated and try to mimic the human brain. They use a very simple modeling of *nerve cells* by functions of the form

$$f : \mathbb{R}^d \to \mathbb{R}, \quad f(o_1, \ldots, o_d) = \sigma\left(w_0 + w_1 \cdot o_1 + \cdots + w_d \cdot o_d\right),$$

where $w_0, \ldots, w_d \in \mathbb{R}$ are the *weights* and $\sigma : \mathbb{R} \to \mathbb{R}$ is the *activation function*.
Traditionally, functions which make a kind of thresholding are used for the activation function, which can be described as follows:

Definition 1.1 *A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called a squashing function if it is nondecreasing and satisfies*

$$\lim_{x \to -\infty} \sigma(x) = 0 \quad \textit{and} \quad \lim_{x \to \infty} \sigma(x) = 1.$$

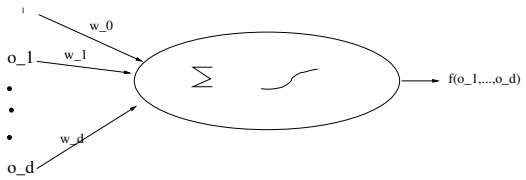

Figure 1: Simple modelling of nerve cell

The simplest example of a squashing function is the *treshold squasher*

$$\sigma(x) = I_{\{x\in[0,\infty)\}},$$

which is not differentiable. A popular example of a differentiable squashing function is the *logistic squasher*

$$\sigma(x) = \frac{1}{1+\exp(-x)}.$$

Nowadays often the so-called *ReLU activation function* (rectified linear unit) is used as activation function:

$$\sigma(x) = \max\{x, 0\}.$$

Using the above functions as building blocks, feedforward neural networks model networks of nerve cells by functions $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f_{\mathbf{w}}(x) = \sum_{k=1}^{r} w_k^{(L)} \cdot f_{\mathbf{w},k}^{(L)}(x) \tag{1}$$

where

$$f_{\mathbf{w},i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} w_{i,j}^{(l-1)} \cdot f_{\mathbf{w},j}^{(l-1)}(x) + w_{i,0}^{(l-1)}\right) \tag{2}$$

$(i \in \{1,\dots,r\}, l \in \{2,\dots,L\})$ and

$$f_{\mathbf{w},i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{i,j}^{(0)} \cdot x^{(j)} + w_{i,0}^{(0)}\right) \tag{3}$$

$(i \in \{1,\dots,r\})$.
Here σ is again the activation function, $r \in \mathbb{N}$ is the *width* of the network and $L \in \mathbb{N}$ is the *depth* of the network. The vector $\mathbf{w} \in \mathbb{R}^K$ consists of all weights

$$w_k^{(L)}, w_{i,j}^{(l-1)},$$

i.e., $\mathbf{w}$ has

$$K = r + (L-1)\cdot r\cdot(r+1) + r\cdot(d+1)$$

many components.

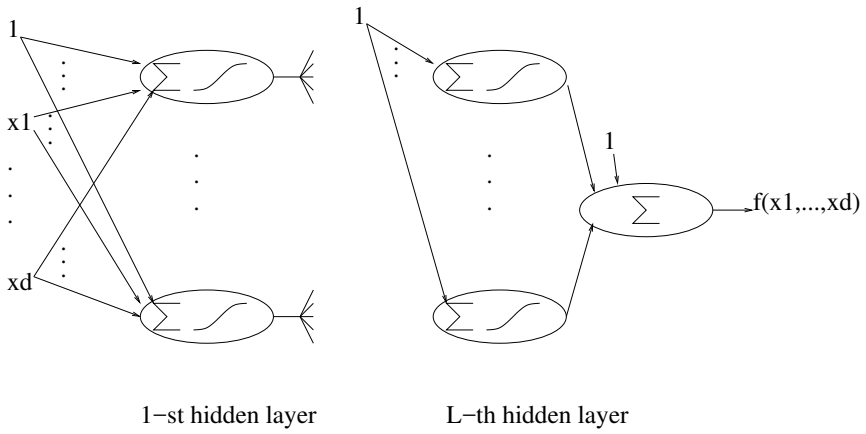

Figure 2: Modelling of network of nerve cells

Definition 1.2 *The space*

$$\mathcal{F}(L, r) = \{f_{\mathbf{w}} \,:\, \mathbf{w} \in \mathbb{R}^K\}$$

of all functions $f_{\mathbf{w}}$ defined by (1)-(3), where the topology (i.e., the structure) and the activation function of the networks are fixed and where the weights vary, is called the space of feedforward neural networks with depth L, width r and activation function σ.

Remark. a) The above function space is useful for regression, where the neural network is used to predict a real number. For pattern recognition and classification, where one of a finite number is to be predicted, the last layer of the network is often modified.
b) The above topology of the networks (L layers of hidden neurons with r neurons in each layer) can be modified, e.g., by using layers with varying widths or by using *sparse* neural networks where not all neurons of one layer have connections to all neurons in the next layer, or by using skip connections between the layers (i.e., connections between layer l and layers $0, \ldots, l-2$).

1.2 Convolutional neural networks

Convolutional networks are applied to input which has some structure. In the sequel we introduce them in connection which images, where the image is described by some $d_1 \times d_2$ dimensional matrix x with values in $[0,1]^{d_1 \times d_2}$ and the entry $x_{i,j}$ at position (i, j) describes the grey scale value of the image at position (i, j).
Convolutional neural networks can be considered as feedforward neural networks where

1. The neurons are arranged in planes with positions corresponding to positions in the image. In each layer there are several neurons for each position in the image, and each neuron is connected only with neurons from the previous layer which correspond to positions "close" to its position.

2. The weights are shared, more precisely: the same weights are used for neurons with correspond to different positions.

3. There are additional pooling and subsampling layers.

We describe this as follows:
Let σ be the activation function, e.g. the ReLU function $\sigma(x) = \max\{x, 0\}$. We next give an example for an convolutional neural network. We start with

$$o^{(0)}_{(i,j),1} = x_{i,j}$$

($(i,j) \in D = \{1,\dots,d_1\} \times \{1,\dots,d_2\}$), and then compute the output of L *convolutional layers* by

$$o^{(l)}_{(i,j),s_2} = \sigma\left(\sum_{s_1=1}^{k_{l-1}} \sum_{\substack{t_1,t_2\in\{1,\dots,M_l\}\\(i+t_1-1,j+t_2-1)\in D}} w^{(l)}_{t_1,t_2,s_1,s_2} o^{(l-1)}_{(i+t_1-1,j+t_2-1),s_1} + w^{(l)}_{s_2}\right)$$

for $(i,j) \in D$, $s_2 \in \{1,\dots,k_l\}$ and $l \in \{1,\dots,L\}$, where $k_0 = 1$, $k_1,\dots,k_{L-1} \in \mathbb{N}$ are the *number of channels* and $M_1,\dots,M_l \in \{1,2,\dots,\min\{d_1,d_2\}\}$ are the *filter sizes*.
Here one may see that weights generating the feature map $o^{(l)}_{(:,:),s_2}$ are shared, which has the advantage that it can reduce the model complexity and the duration of the networks' training.
Finally we apply a *global max-pooling layer*:

$$f_{\mathbf{w},\mathbf{w}_{bias},\mathbf{w}_{out}}(\mathbf{x}) = \max\left\{\sum_{s_2=1}^{k_L} w_{s_2} \cdot o^{(L)}_{(i,j),s_2} : i \in \{1,\dots,d_1 - B + 1\}, j \in \{1,\dots,d_2 - B + 1\}\right\},$$

where $B \in \{1,2,\dots,\min\{d_1,d_2\}\}$ is an output bound and where

$$\mathbf{w} = \left(w^{(l)}_{t_1,t_2,s_1,s_2}\right)_{1\le t_1,t_2\le M_l, s_1\in\{1,\dots,k_{l-1}\}, s_2\in\{1,\dots,k_l\}, l\in\{1,\dots,L\}},$$

$$\mathbf{w}_{bias} = (w^{(l)}_{s_2})_{s_2\in\{1,\dots,k_r\}, l\in\{1,\dots,L\}}$$

and

$$\mathbf{w}_{out} = (w_s)_{s\in\{1,\dots,k_L\}}$$

are the weights of our convolutional network. To the output of this convolutional network we can then apply a feedforward neural network.
Alternatively we could also insert a *subsampling layer* (or a local max-pooling layer) at some point in the computation, which reduces the resultion of the image, e.g., (in case of a subsampling layer) by choosing $r \in \{1,\dots,\min\{d_1,d_2\}\}$ and by setting

$$o^{(l)}_{(i,j),s_2} = o^{(l+1)}_{(1+(i-1)\cdot r,1+(j-1)\cdot r),s_2}$$

for

$$(i,j) \in \left\{1,\dots,\lceil\frac{d_1}{r}\rceil\right\} \times \left\{1,\dots,\lceil\frac{d_2}{r}\rceil\right\} =: D'$$

at some layer l and by using D' instead of D after this layer.
There exists various modifications of the above topology of a convolutional neural network.

1.3 Transformer networks

Transformer networks are applied to data which consists of sequences of data points. We describe them next in the context of text data, i.e., our data is a sequence of words (which we want, e.g., to translate in another language or which we want to classify, e.g., as hate speech or not hate speech).
The first step is to transform the sequence of words in a sequences of real numbers, where each word is decoded by a vector in $\mathbb{R}^d$ *(word to vector conversion)*. Here the text is first tokenized, then shared subword units are learned by starting with a symbol vocabulary in which successively the most frequent occuring pairs of symbols within the tokens are merged and are replaced by some newly introduced symbol until the final vocabulary size (e.g., 32,000) is reached. Then each symbol in the vocabulary is assigned to a vector in $\mathbb{R}^{d_{model}}$ (either using values from some pretraining or using random values), and the text to be considered is replaced by the sequence of vectors corresponding to its symbols. Here often some additional coding of the position in the text (consisting of values from sinus and cosinus) are added.
From now on we assume that each data point is of the form

$$x = (x_1,\dots,x_l) \in \mathbb{R}^{d_{model}\times l}.$$

In principle, we could try to apply a feedforward neural network to such data. But since $d_{model}\cdot l$ might be rather large, this network will need too many parameters. One solution is to use *recurrent neural network*, which see successively x_1, x_2, ..., x_l and use some of the outputs they have computed just after seeing x_i as additional inputs for the computation of its output for x_{i+1}. So these networks have *recurrent connections* as in Figure 3.
The problem with this approach is, that in text data there might by long ranging dependencies between the words. Consider the following two examples of the occurence of the word "bank":

1. I do not like this bank. The reason is that each time I walk into it, there is a long queue at the counter.

2. I do not like this bank. The reason is that each time I sit on it my back hurts.

Here you cannot determine the meaning of "bank" before you have seen words which come long after the word "bank".

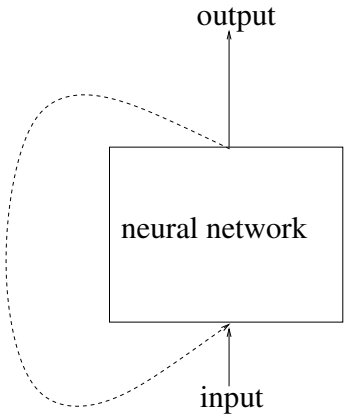

Figure 3: Example of a recurrent network.

Instead of feedforward neural layers the transformer networks use another mechanism to compute their output, the so-called maximal attention layers.
Starting with the above new representation

$$z_0 = (z_{0,1}, \ldots, z_{0,l}) \in \mathbb{R}^{d_{model} \times l}$$

of the original input the transformer network computes successively representations

$$z_r = (z_{r,1}, \ldots, z_{r,l}) \in \mathbb{R}^{d_{model} \times l} \tag{4}$$

of the input for $r = 1, \ldots, N$, and applies a feedforward neural network to z_N. Here z_r is the representation of the input in level r. It depends on l parts which correspond to x_1, $\ldots$, x_l. And N is the number of pairs of attention and pointwise feedforward layers of our transformer encoder.
Given z_{r-1} for some $r \in \{1, \ldots, N\}$ we compute z_r by applying first a *multi-head attention* and afterwards a *pointwise feedforward neural network with one hidden layer*. Both times we will use an additional residual connection.
The computation of the multi-head attention depends on matrices

$$W_{query,r,s}, W_{key,r,s} \in \mathbb{R}^{d_{key} \times d_{model}} \quad \text{and} \quad W_{value,r,s} \in \mathbb{R}^{d_v \times d_{model}} \quad (s = 1, \ldots, h),$$

where $h \in \mathbb{N}$ is the number of attention heads which we compute in parallel, where $d_{key} \in \mathbb{N}$ is the dimension of the queries and the keys, and where $d_v = d_{model}/h$ is the dimension of the values. Here it is assumed that d_{model}/h is a natural number and each of the h attention heads will be used to compute a new part of length d_v of the representation $z_{r,i}$ of x_i for $i = 1, \ldots, l$. We use the above matrices to compute for each component $z_{r-1,i}$ of z_{r-1} (i.e., for each representation of x_i at level $r-1$ $(i = 1, \ldots, l)$) corresponding *queries*

$$q_{r-1,s,i} = W_{query,r,s} \cdot z_{r-1,i},$$

keys

$$k_{r-1,s,i} = W_{key,r,s} \cdot z_{r-1,i}$$

and *values*

$$v_{r-1,s,i} = W_{value,r,s} \cdot z_{r-1,i}$$

$(s \in \{1,\dots,h\}, i \in \{1,\dots,l\})$. Then the so-called *attention* between the component i of z_{r-1} and the component j of z_{r-1} (i.e., between the representations of x_i and x_j at level $r-1$) is defined as the scalar product

$$< q_{r-1,s,i}, k_{r-1,s,j} > \tag{5}$$

of the corresponding query and key, and the index $\hat{j}_{r-1,s,i}$ for which the maximal value occurs, i.e.,

$$\hat{j}_{r-1,s,i} = \arg \max_{j \in \{1,\dots,l\}} < q_{r-1,s,i}, k_{r-1,s,j} >,$$

is determined. The value corresponding to this index is multiplied with the maximal attention in (5) in order to define

$$\begin{aligned} \bar{y}_{r,s,i} &= v_{r-1,s,\hat{j}_{r-1,s,i}} \cdot \max_{j \in \{1,\dots,l\}} < q_{r-1,s,i}, k_{r-1,s,j} > \\ &= v_{r-1,s,\hat{j}_{r-1,s,i}} \cdot < q_{r-1,s,i}, k_{r-1,s,\hat{j}_{r-1,s,i}} > \end{aligned}$$

$(s \in \{1,\dots,h\}, i \in \{1,\dots,l\})$. Using a residual connection we compute the output of the multi-head attention by

$$y_r = z_{r-1} + (\bar{y}_{r,1}, \dots, \bar{y}_{r,l}) \tag{6}$$

where

$$\bar{y}_{r,i} = (\bar{y}_{r,1,i}, \dots, \bar{y}_{r,h,i}) \in \mathbb{R}^{d_v \cdot h} = \mathbb{R}^{d_{model}} \quad (i \in \{1,\dots,l\}).$$

Here $y_r \in \mathbb{R}^{d_{model} \times l}$ has the same dimension as z_{r-1}.
The output of the pointwise feedforward neural network depends on parameters

$$W_{r,1} \in \mathbb{R}^{d_{ff} \times d_{model}}, b_{r,1} \in \mathbb{R}^{d_{ff}}, W_{r,2} \in \mathbb{R}^{d_{model} \times d_{ff}}, b_{r,2} \in \mathbb{R}^{d_{model}},$$

which describe the weights in a feedforward neural network with one hidden layer and $d_{ff} \in \mathbb{N}$ hidden neurons. This feedforward neural network is applied to each component of (6) (which is analogous to a convolutionary neural network), i.e., to each representation of $x_1, \dots, x_l$ computed up to this point on level r, and computes

$$z_{r,i} = y_{r,i} + W_{r,2} \cdot \sigma\left(W_{r,1} \cdot y_{r,i} + b_{r,1}\right) + b_{r,2} \quad (i \in \{1,\dots,l\}),$$

where we use again a residual connection. Here

$$\sigma(x) = \max\{x, 0\}$$

is the ReLU activation function, which is applied to a vector by applying it to each component of the vector separately. After computing $z_{r,i}$ $(i \in \{1,\dots,l\})$ we define z_r by (4).

Given the output z_N of the sequence of N multi-head attention and pointwise feedforward layers, we can apply a feedforward neural network to this output to compute the output of our transformer network.
In applications the maximal attention is defined by using the so–called *softmax* function. Here we set for nonnegative real numbers $u_1, \ldots, u_K$

$$softmax(u_1, \ldots, u_K) = \left(\frac{e^{u_1}}{\sum_{k=1}^K e^{u_k}}, \ldots, \frac{e^{u_K}}{\sum_{k=1}^K e^{u_k}} \right),$$

which is a vector in $[0,1]^K$ which ideally has a one in the maximal component of $(u_1, \ldots, u_K)$ and zeros in all other components. Using this function one defines the maximal attention by

$$\bar{y}_{r,s,i} \quad = \quad \frac{e^{<q_{r-1,s,i}, k_{r-1,s,1}>}}{\sum_{j=1}^l e^{<q_{r-1,s,i}, k_{r-1,s,j}>}} \cdot v_{r-1,s,1} + \cdots + \frac{e^{<q_{r-1,s,i}, k_{r-1,s,l}>}}{\sum_{j=1}^l e^{<q_{r-1,s,i}, k_{r-1,s,j}>}} \cdot v_{r-1,s,l}.$$

Above we have described a *Transformer encoder*. In applications like machine translation an *encoder-decoder structure* is used.

1.4 Learning of deep (neural) networks

In statistical applications of deep learning the aim is to fit a deep (neural) network

$$f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$$

to data

$$(x_1, y_1), \ldots, (x_n, y_n),$$

where the data points can be, e.g., points from $\mathbb{R}^d \times \mathbb{R}$ or from $\mathbb{R}^d \times \{-1, 1\}$. Here $f_{\mathbf{w}}$ is computed by

$$f_{\mathbf{w}}(x) = \sum_{k=1}^r w_{1,k}^{(L)} \cdot o_k^{(L)}, \tag{7}$$

$$o_k^{(l)} = \sigma(a_k^{(l)}) \quad \text{for } l \in \{1, \ldots, L\}, \tag{8}$$

$$a_i^{(l)} = \sum_{j=1}^r w_{i,j}^{(l-1)} \cdot o_j^{(l-1)} + w_{i,0}^{(l-1)} \tag{9}$$

for $l \in \{2, \ldots, L\}$ and

$$a_i^{(1)} = \sum_{j=1}^d w_{i,j}^{(0)} \cdot x^{(j)} + w_{i,0}^{(0)}. \tag{10}$$

From (7)-(10) we see that the input is propagated through the network in order to compute the function value, i.e., we compute successively

$$\begin{aligned}
&a_1^{(1)}, \ldots, a_d^{(1)} \\
&o_1^{(1)}, \ldots, o_d^{(1)} \\
&a_1^{(2)}, \ldots, a_r^{(2)} \\
&o_1^{(2)}, \ldots, o_r^{(2)} \\
&\vdots \\
&a_1^{(L)}, \ldots, a_r^{(L)} \\
&o_1^{(L)}, \ldots, o_r^{(L)} \\
&f_{\mathbf{w}}(x).
\end{aligned}$$

In order to compute the weight vector $\mathbf{w}$ of our network, we first choose a so–called *loss function*

$$l(y, z) \geq 0,$$

e.g.,

$$l(y, z) = \frac{1}{2} \cdot (y - z)^2$$

(*least squares loss*) for $y, z \in \mathbb{R}$ or

$$l(y, z) = \log(1 + \exp(-y \cdot z))$$

(*logistic loss*) for $y \in \{-1, 1\}$, $z \in \mathbb{R}$. Then we try to choose $\mathbf{w}$ such that the *empirical risk*

$$\hat{r}(w) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f_{\mathbf{w}}(x_i)) \tag{11}$$

is small.
Since $l(y, f_{\mathbf{w}}(x))$ depends nonlinearly on $\mathbf{w}$, we use *gradient descent* to achieve this: We choose a starting value

$$\mathbf{w}^{(0)}$$

(e.g., randomly from some proper distribution), and set

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda \cdot \nabla_{\mathbf{w}} \left(\hat{r}(\mathbf{w}^{(t)}) \right)$$

for $t = 0, 1, \ldots, T-1$, where $\lambda > 0$ is the *stepsize* and $T \in \mathbb{N}$ is the *number of gradient descent steps*.
Here the gradient $\nabla_{\mathbf{w}} \left(\hat{r}(\mathbf{w}) \right)$ is the vector of all partial derivatives

$$\frac{\partial}{\partial w_{i,j}^{(l)}} \hat{r}(\mathbf{w}), \tag{12}$$

so in order to compute the gradient we need to be able to compute the partial derivatives in (12).

This is possible by a simple algorithm (so-called *backpropagation*), which we derive next by an application of the chain rule.

We have

$$\frac{\partial}{\partial w_{i,j}^{(l)}}\hat{r}(\mathbf{w}) = \frac{1}{n}\sum_{s=1}^{n}\frac{\partial}{\partial z}l(y_s, f_{\mathbf{w}}(x_s)) \cdot \frac{\partial}{\partial w_{i,j}^{(l)}}f_{\mathbf{w}}(x_s),$$

where

$$\frac{\partial}{\partial w_{1,j}^{(L)}}f_{\mathbf{w}}(x_s) = \frac{\partial}{\partial w_{1,j}^{(L)}}\left(\sum_{k=1}^{r} w_{1,k}^{(L)} \cdot o_k^{(L)}\right) = o_j^{(L)},$$

and for $l < L$

$$\frac{\partial}{\partial w_{i,j}^{(l)}}f_{\mathbf{w}}(x_s) = \sum_{s=1}^{r} w_{1,s}^{(L)} \cdot \sigma'(a_s^{(L)}) \cdot \frac{\partial}{\partial w_{i,j}^{(l)}}(a_s^{(L)}).$$

Furthermore, for $l \in \{2, \ldots, L\}$ we have

$$\frac{\partial}{\partial w_{k,j}^{(l-1)}}(a_k^{(l)}) = o_j^{(l-1)}, \quad \frac{\partial}{\partial w_{i,j}^{(l-1)}}(a_k^{(l)}) = 0 \quad (i \neq k)$$

(with $o_0^{(l-1)} = 1$), and in case $m < l - 1$

$$\frac{\partial}{\partial w_{i,j}^{(m)}}(a_k^{(l)}) = \sum_{s=1}^{r} w_{k,s}^{(l-1)} \cdot \sigma'(a_s^{(l-1)}) \cdot \frac{\partial}{\partial w_{i,j}^{(m)}}(a_s^{(l-1)}).$$

In addition, it holds

$$\frac{\partial}{\partial w_{i,j}^{(0)}}(a_i^{(1)}) = x^{(j)}, \quad \frac{\partial}{\partial w_{i,j}^{(0)}}(a_k^{(1)}) = 0 \quad (i \neq k)$$

(with $x^{(0)} = 1$).

In case of $l(y, z) = \frac{1}{2} \cdot (y - z)^2$ we have

$$\frac{\partial}{\partial z}l(y_s, f_{\mathbf{w}}(x_s)) = (y_s - f_{\mathbf{w}}(x_s)) \cdot (-1),$$

and we see that with the formulas above the so–called residual error

$$y_s - f_{\mathbf{w}}(x_s)$$

of the neural network is propagated back through the network during the computation of the gradient. More precisely, if we recursively compute

$$\delta_1^{(L+1)} = f_{\mathbf{w}}(x_s) - y_s,$$

$$\delta_i^{(L)} = \sigma'(a_i^{(L)}) \cdot w_{1,i}^{(L)} \cdot \delta_1^{(L+1)} \quad (i = 1, \ldots, r)$$

and

$$\delta_i^{(l)} = \sigma'(a_i^{(l)}) \cdot \sum_{j=1}^{r} w_{j,i}^{(l)} \cdot \delta_j^{(l+1)} \quad (i = 1, \ldots, r, \quad l = 1, \ldots, L-1)$$

then we have

$$(f_{\mathbf{w}}(x_s) - y_s) \cdot \frac{\partial f_{\mathbf{w}}(x_s)}{\partial w_{i,j}^{(l)}} = \delta_i^{(l+1)} \cdot o_j^{(l)}.$$

Here the above formulas can be either verified by the formulas derived above, or directly derived from the chain rule provided we set

$$\delta_j^{(l)} = \frac{\partial f_{\mathbf{w}}(x_s)}{\partial a_j^{(l)}}.$$

To see this observe

$$\frac{\partial f_{\mathbf{w}}(x_s)}{\partial w_{i,j}^{(l)}} = \frac{\partial f_{\mathbf{w}}(x_s)}{\partial a_i^{(l+1)}} \cdot \frac{\partial a_i^{(l+1)}}{\partial w_{i,j}^{(l)}}$$

and

$$\frac{\partial f_{\mathbf{w}}(x_s)}{\partial a_i^{(m)}} = \sum_{k=1}^{r} \frac{\partial f_{\mathbf{w}}(x_s)}{\partial a_k^{(m+1)}} \cdot \frac{\partial a_k^{(m+1)}}{\partial a_i^{(m)}}$$

for $m < L$.

2 Neural network approximation

2.1 Introduction

In this chapter we investigate how well *smooth* functions $f : \mathbb{R}^d \to \mathbb{R}$ can be approximated by neural networks. We consider only ReLU networks, neural networks with squashing functions as activation functions will be considered in the practicing course.
We use the following definition in order to describe what are the smooth functions which we want to approximate by neural networks:

Definition 2.1 *Let $p = k+\beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k f}{\partial^{\alpha_1} x^{(1)} \ldots \partial^{\alpha_d} x^{(d)}}$ exists and satisfies*

$$\left| \frac{\partial^k f}{\partial^{\alpha_1} x^{(1)} \ldots \partial^{\alpha_d} x^{(d)}}(x) - \frac{\partial^k f}{\partial^{\alpha_1} x^{(1)} \ldots \partial^{\alpha_d} x^{(d)}}(z) \right| \leq C \cdot \|x - z\|^\beta$$

for all $x, z \in \mathbb{R}^d$, where $\mathbb{N}_0$ is the set of non-negative integers.

So (p, C)–smooth functions are functions whose derivatives of order k are Hölder continuous with exponent β and Hölder constant C. In particular, (p, C)–smooth functions with $p \leq 1$ are Hölder continuous with Hölder exponent p.
The aim in the sequel is to derive for given spaces $\mathcal{F}$ of neural networks upper bounds on

$$d(g, \mathcal{F}, \|\cdot\|_{\infty,[-A,A]^d}) = \inf_{f \in \mathcal{F}} \|g - f\|_{\infty,[-A,A]^d}$$

for (p, C)–smooth functions $g : \mathbb{R}^d \to \mathbb{R}$. Here we are particularly interested how the derived upper bound depends on the *number of nonzero weights* in the space $\mathcal{F}$ of neural networks.
In order to understand what we can expect as a result we first derive a result for a space of piecewise polynomials. To do this, we use the next lemma.

Lemma 2.1 *Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)–smooth function, let $x_0 \in \mathbb{R}^d$ and let p_k be the Taylor polynomial of f of total degree k around x_0, i.e.,*

$$p_k(x) = \sum_{\substack{j_1,\ldots,j_d \in \{0,1,\ldots,k\}, \\ j_1+\cdots+j_d \leq k}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \ldots \partial^{j_d} x^{(d)}}(x_0) \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d}.$$

Then we have for any $x \in \mathbb{R}^d$

$$|f(x) - p_k(x)| \leq c_1 \cdot C \cdot \|x - x_0\|^p$$

for some constant $c_1 \in \mathbb{R}$ depending only on k and on d.

Proof. We start the proof with a repition of some basic facts about Taylor polynomials. Let $f : \mathbb{R} \to \mathbb{R}$ be $(k+1)$–times differentiable function, then integration by parts shows the standard integral form of the remainder of the Taylor polynomial:

$$f(x) = f(x_0) + \frac{f^{(1)}(x_0)}{1!} \cdot (x - x_0)^1 + \cdots + \frac{f^{(k)}(x_0)}{k!} \cdot (x - x_0)^k + \int_{x_0}^{x} \frac{f^{(k+1)}(t)}{k!} \cdot (x - t)^k dt.$$

Next let $f : \mathbb{R}^d \to \mathbb{R}$ be (p, C)–smooth and define

$$h(s) = f(x_0 + s \cdot (x - x_0)) \quad (s \in \mathbb{R}).$$

Then the above formula implies

$$f(x) = h(1) = h(0) + \frac{h^{(1)}(0)}{1!} \cdot (1 - 0)^1 + \cdots + \frac{h^{(k)}(0)}{k!} \cdot (1 - 0)^k + \int_0^1 \frac{h^{(k+1)}(t)}{k!} \cdot (1 - t)^k dt.$$

Using the chain rule we get

$$\begin{aligned}
\frac{h^{(k)}(s)}{k!} &= \frac{1}{k!} \cdot \sum_{j_1=1}^{d} \cdots \sum_{j_k=1}^{d} \frac{\partial^k f}{\partial x^{(j_1)} \dots \partial x^{(j_k)}}(x_0 + s \cdot (x - x_0)) \cdot (x^{(j_1)} - x_0^{(j_1)}) \cdots (x^{(j_k)} - x_0^{(j_k)}) \\
&= \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\}, \\ j_1+\cdots+j_d=k}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0 + s \cdot (x - x_0)) \\
&\qquad \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d}.
\end{aligned}$$

Combining this with the previous result we get the integral form of the remainder of the multivariate Taylor polynomial of order $k - 1$:

$$\begin{aligned}
f(x) &= \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k-1\}, \\ j_1+\cdots+j_d \le k-1}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0) \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d} \\
&\quad + \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\}, \\ j_1+\cdots+j_d=k}} \frac{k}{j_1! \cdots j_d!} \cdot \int_0^1 (1-t)^{k-1} \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0 + t \cdot (x - x_0))\, dt \\
&\qquad \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d}.
\end{aligned}$$

Now we start our proof. The definition of p_k and the above integral form of the remainder of a Taylor series imply

$$
\begin{aligned}
&f(x) - p_k(x) \\
&= f(x) - p_{k-1}(x) - \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\},\\ j_1+\cdots+j_d=k}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0) \\
&\qquad \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d} \\
&= \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\},\\ j_1+\cdots+j_d=k}} \frac{k}{j_1! \cdots j_d!} \cdot \int_0^1 (1-t)^{k-1} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0 + t \cdot (x - x_0))\, dt \\
&\qquad \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d} \\
&\quad - \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\},\\ j_1+\cdots+j_d=k}} \frac{k}{j_1! \cdots j_d!} \cdot \int_0^1 (1-t)^{k-1} \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0)\, dt \\
&\qquad \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d} \\
&= \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\},\\ j_1+\cdots+j_d=k}} \frac{k}{j_1! \cdots j_d!} \cdot (x^{(1)} - x_0^{(1)})^{j_1} \cdots (x^{(d)} - x_0^{(d)})^{j_d} \\
&\quad \cdot \int_0^1 (1-t)^{k-1} \left(\frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0 + t \cdot (x - x_0)) - \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(x_0) \right) dt.
\end{aligned}
$$

Using the triangle inequality and the (p, C)–smoothness of f we conclude

$$
\begin{aligned}
&|f(x) - p_k(x)| \\
&\leq \sum_{\substack{j_1,\dots,j_d \in \{0,1,\dots,k\},\\ j_1+\cdots+j_d=k}} \frac{k}{j_1! \cdots j_d!} \cdot \|x - x_0\|^k \cdot \int_0^1 (1-t)^{k-1} \cdot C \cdot t^\beta \cdot \|x - x_0\|^\beta \, dt \\
&\leq c_1 \cdot C \cdot \|x - x_0\|^{k+\beta},
\end{aligned}
$$

which completes the proof. □

Let $A > 0$, $K \in \mathbb{N}$ and consider a partition of $[-A, A]^d$ in K^d cubes of side length $(2 \cdot A)/K$. Let $\mathcal{F}$ be the set of all functions which are on each set of this partition equal to a polynomial of total degree k or less. Let g be an arbitrary (p, C)–smooth function. By choosing f on each set in the partition as the Taylor polynomial to g of Lemma 2.1 with x_0 arbitrary chosen from the considered set, we get by Lemma 2.1

$$
\|g - f\|_{\infty,[-A,A]^d} \leq c_1 \cdot C \cdot \left(\sqrt{d} \cdot \frac{2A}{K} \right)^p,
$$

so we see that the above space of piecewise polynomials (which has $c_2 \cdot K^d$ many parameters) satisfies

$$
d(g, \mathcal{F}, \|\cdot\|_{\infty,[-A,A]^d}) \leq c_3 \cdot \frac{1}{K^p}
$$

for any (p, C)–smooth function. This implies, that with a suitable space of piecewise polynomials with K parameters we get

$$
d(g, \mathcal{F}, \|\cdot\|_{\infty,[-A,A]^d}) = O\left(K^{-\frac{p}{d}} \right)
$$

for any (p, C)–smooth function $g : \mathbb{R}^d \to \mathbb{R}$.
In the sequel we will show a similar result for deep neural networks.

2.2 Approximation power of deep neural networks

Let $A > 0$ be fixed. We define a local convex combination of Taylor polynomials, which we will later approximate by deep neural networks.
For $K \in \mathbb{N}$ and $\mathbf{i} = (i^{(1)}, \dots, i^{(d)}) \in \{0, \dots, K\}^d$ set

$$x_{\mathbf{i}} = \left(-A + i^{(1)} \cdot \frac{2A}{K}, \dots, -A + i^{(d)} \cdot \frac{2A}{K} \right)$$

and let

$$\{\mathbf{i}_1, \dots, \mathbf{i}_{(K+1)^d}\} = \{0, \dots, K\}^d.$$

For $k \in \{1, \dots, (K+1)^d\}$ let

$$p_{\mathbf{i}_k}(x) = \sum_{\substack{j_1, \dots, j_d \in \{0, \dots, q\} \\ j_1 + \dots + j_d \le q}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1 + \dots + j_d} g}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_{\mathbf{i}_k}) \cdot (x^{(1)} - x^{(1)}_{\mathbf{i}_k})^{j_1} \cdots (x^{(d)} - x^{(d)}_{\mathbf{i}_k})^{j_d}$$

be the the Taylor polynomial of g with order q around $x_{\mathbf{i}_k}$ and set

$$\begin{aligned} P(x) &= \sum_{k=1}^{(K+1)^d} p_{\mathbf{i}_k}(x) \cdot \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}| \right)_+ \\ &= \sum_{k=1}^{(K+1)^d} \sum_{\substack{j_1, \dots, j_d \in \{0, \dots, q\} \\ j_1 + \dots + j_d \le q}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1 + \dots + j_d} g}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_{\mathbf{i}_k}) \cdot \prod_{i=1}^{d} (x^{(i)} - x^{(i)}_{\mathbf{i}_k})^{j_i} \\ &\qquad \cdot \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}| \right)_+, \end{aligned} \tag{13}$$

where $z_+ = \max\{z, 0\}$ $(z \in \mathbb{R})$.
Because of

$$\prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}| \right)_+ \ge 0 \quad \text{and} \quad \sum_{k=1}^{(K+1)^d} \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}| \right)_+ = 1$$

for $x \in [-A, A]^d$ (where the equality holds because of

$$\begin{aligned} &\sum_{k=1}^{(K+1)^d} \prod_{j=1}^{d} \left(1 - \frac{K}{2a} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}| \right)_+ \\ &= \sum_{k=0}^{K} \left(1 - \frac{K}{2A} \cdot |x^{(1)} - (-A + k \cdot \frac{2A}{K})| \right)_+ \cdot \ldots \cdot \sum_{k=0}^{K} \left(1 - \frac{K}{2A} \cdot |x^{(d)} - (-A + k \cdot \frac{2A}{K})| \right)_+ \end{aligned}$$

and

$$\sum_{k=0}^{K}\left(1-\frac{K}{2A}\cdot|u-(-A+k\cdot\frac{2A}{K})|\right)_+=1$$

for all $u\in[-A,A]$), and

$$\prod_{j=1}^{d}\left(1-\frac{K}{2A}\cdot|x^{(j)}-x_{\mathbf{i}_k}^{(j)}|\right)_+=0$$

if $\|x-x_{\mathbf{i}_k}\|_\infty\geq 2A/K$, $P(x)$ is a local convex combination of Taylor polynomials of m. Consequently we can conclude from Lemma 2.1 that for any (p,C)–smooth function $g:\mathbb{R}^d\to\mathbb{R}$ (where $p=q+\beta$ for some $q\in\mathbb{N}_0$ and $0<\beta\leq 1$) and for any $x\in[-A,A]^d$ we have

$$\begin{aligned}
&|P(x)-g(x)|\\
&=\left|\sum_{k=1}^{(K+1)^d}(p_{\mathbf{i}_k}(x)-g(x))\prod_{j=1}^{d}\left(1-\frac{K}{2A}\cdot|x^{(j)}-x_{\mathbf{i}_k}^{(j)}|\right)_+\right|\\
&\leq\sum_{k=1}^{(K+1)^d}|p_{\mathbf{i}_k}(x)-g(x)|\prod_{j=1}^{d}\left(1-\frac{K}{2A}\cdot|x^{(j)}-x_{\mathbf{i}_k}^{(j)}|\right)_+\\
&=\sum_{k=1,\ldots,(K+1)^d,\|x-x_{\mathbf{i}_k}\|_\infty<2A/K}|p_{\mathbf{i}_k}(x)-g(x)|\prod_{j=1}^{d}\left(1-\frac{K}{2A}\cdot|x^{(j)}-x_{\mathbf{i}_k}^{(j)}|\right)_+\\
&\leq c_1\cdot\frac{1}{K^p}\cdot\sum_{k=1,\ldots,(K+1)^d,\|x-x_{\mathbf{i}_k}\|_\infty<2A/K}\prod_{j=1}^{d}\left(1-\frac{K}{2A}\cdot|x^{(j)}-x_{\mathbf{i}_k}^{(j)}|\right)_+\leq c_1\cdot\frac{1}{K^p}.
\end{aligned}$$

In the sequel we derive a neural network with ReLU activation function which approximates (13).

Our starting point is the following approximation of the square function by a deep ReLU network.

Lemma 2.2 *(Yarotsky (2017)) Let $\sigma:\mathbb{R}\to\mathbb{R}$ be the ReLU activation function $\sigma(x)=\max\{x,0\}$. Then for any $R\in\mathbb{N}$ and any $a\geq 1$ a neural network*

$$\hat{f}_{sq}\in\mathcal{F}(R,9)$$

exists such that

$$\left|\hat{f}_{sq}(x)-x^2\right|\leq a^2\cdot 4^{-R}$$

holds for $x\in[-a,a]$.

Proof. We consider the "tooth" function $g:[0,1]\to[0,1]$

$$g(x)=\begin{cases}2x & ,x\leq\frac{1}{2}\\ 2\cdot(1-x) & ,x>\frac{1}{2}\end{cases}$$

and the iterated function

$$g_s(x) = \underbrace{g \circ g \circ \cdots \circ g}_{s \text{ times}}(x).$$

In a *first step of the proof* we show by induction that

$$g_s(x) = \begin{cases} 2^s\left(x - \frac{2k}{2^s}\right) & , x \in \left[\frac{2k}{2^s}, \frac{2k+1}{2^s}\right], k = 0, 1, \ldots, 2^{s-1} - 1 \\ 2^s\left(\frac{2k}{2^s} - x\right) & , x \in \left[\frac{2k-1}{2^s}, \frac{2k}{2^s}\right], k = 1, 2, \ldots, 2^{s-1} \end{cases}.$$

For $s = 1$ this follows directly from the definition of g and g_1. For the induction step we remark that $(g_s \circ g)(x) = g_s(2x)$ whenever $x \in [0, \frac{1}{2}]$ and that $g(x) = g(1 - x)$. This combined with the symmetry of g_s (by the inductive hypothesis) implies that for every $x \in [0, \frac{1}{2}]$

$$\begin{aligned} g_{s+1}(x) &= g_s(g(x)) = g_s(2x) = g_s(1 - 2x) = g_s\left(2 \cdot (\frac{1}{2} - x)\right) \\ &= g_s\left(g\left(\frac{1}{2} - x\right)\right) = g_s\left(g\left(x + \frac{1}{2}\right)\right) = g_{s+1}\left(x + \frac{1}{2}\right). \end{aligned}$$

Consequently it suffices to consider $x \in [0, \frac{1}{2}]$ which means

$$(g_s \circ g)(x) = g_s(2x)$$

and together with the inductive hypothesis we have

$$\begin{aligned} (g_s \circ g)(x) &= \begin{cases} 2^s \cdot (2x - \frac{2k}{2^s}) & , 2x \in [\frac{2k}{2^s}, \frac{2k+1}{2^s}], k = 0, 1, \ldots, 2^{s-1} - 1 \\ 2^s \cdot (\frac{2k}{2^s} - 2x) & , 2x \in [\frac{2k-1}{2^s}, \frac{2k}{2^s}], k = 1, 2, \ldots, 2^{s-1} \end{cases} \\ &= \begin{cases} 2^{s+1} \cdot (x - \frac{2k}{2^{s+1}}) & , x \in [\frac{2k}{2^{s+1}}, \frac{2k+1}{2^{s+1}}], k = 0, 1, \ldots, 2^s - 1 \\ 2^{s+1} \cdot (\frac{2k}{2^{s+1}} - x) & , x \in [\frac{2k-1}{2^{s+1}}, \frac{2k}{2^{s+1}}], k = 1, 2, \ldots, 2^s, \end{cases} \end{aligned}$$

which shows the assertion.

In a *second step of the proof* we show that the function $f(x) = x^2$, $x \in [0, 1]$ can be approximated by linear combinations of functions g_s. Let S_R be a piecewise linear interpolation of f with $2^R + 1$ uniformly distributed breakpoints $\frac{k}{2^R}$, $k = 0, \ldots, 2^R$

$$S_R\left(\frac{k}{2^R}\right) = \left(\frac{k}{2^R}\right)^2.$$

To determine the error of that piecewise linear interpolation we define the auxiliary function

$$F(z) = f(z) - S_R(z) + \frac{S_R(x) - f(x)}{(x - \frac{k}{2^R})(x - \frac{k+1}{2^R})} \cdot (z - \frac{k}{2^R})(z - \frac{k+1}{2^R})$$

for $x \in [\frac{k}{2^R}, \frac{k+1}{2^R}]$ and $k = 0, \ldots, 2^R - 1$.
We note that $F(\frac{k}{2^R}) = 0$, $F(\frac{k+1}{2^R}) = 0$ and $F(x) = 0$. According to Rolle's theorem, there

must be a point z_1, where $\frac{k}{2^R} < z_1 < x$ and $F'(z_1) = 0$ and there must be a point z_2, where $x < z_2 < \frac{k+1}{2^R}$ and $F'(z_2) = 0$. Using Rolle's theorem again, there must be a point η where $z_1 < \eta < z_2$ and $F''(\eta) = 0$. Thus we get for any $x \in [\frac{k}{2^R}, \frac{k+1}{2^R}]$

$$0 = F''(\eta) = f''(\eta) + \frac{S_r(x) - f(x)}{(x - \frac{k}{2^R}) \cdot (x - \frac{k+1}{2^R})},$$

which implies

$$\begin{aligned} |f(x) - S_R(x)| &= \left| -\frac{f''(\eta)}{2} \cdot (x - \frac{k}{2^R})(x - \frac{k+1}{2^R}) \right| \\ &\leq \left| (x - \frac{k}{2^R})(x - \frac{k+1}{2^R}) \right| \leq 2^{-2R-2}, \end{aligned}$$

where the last inequality follows since the maximum of

$$h(x) := (x - \frac{k}{2^R})(\frac{k+1}{2^R} - x)$$

is given by $h\left(\frac{k}{2^R} + \frac{1}{2} \cdot \frac{1}{2^R}\right)$.
Furthermore refining the interpolation from S_{R-1} to S_R amounts to adjusting it by a function proportional to a sawtooth function:

$$S_{R-1}(x) - S_R(x) = \frac{g_R(x)}{2^{2R}}.$$

To see this, consider let $k \in \{0, 1, \ldots, 2^{R-1} - 1\}$ and assume $x \in [k/2^{R-1}, (k+1)/2^{R-1}]$. Since $S_{R-1}(x) - S_R(x)$ and $g_R(x)$ are on both intervals $[k/2^{R-1}, (k + 1/2)/2^{R-1}]$ and $[(k + 1/2)/2^{R-1}, (k+1)/2^{R-1}]$ linear, vanish at $k/2^{R-1}$ and $(k+1)/2^{R-1}$ and

$$g_R((k + 1/2)/2^{R-1}) = 1,$$

we have for all $x \in [k/2^{R-1}, (k+1)/2^{R-1}]$:

$$S_{R-1}(x) - S_R(x) = (S_{R-1}((k + 1/2)/2^{R-1}) - S_R((k + 1/2)/2^{R-1})) \cdot g_R(x).$$

By definition of S_{R-1} and S_R we have

$$S_{R-1}((k + 1/2)/2^{R-1}) = \frac{(k/2^{R-1})^2 + ((k+1)/2^{R-1})^2}{2}$$

and

$$S_R((k + 1/2)/2^{R-1})) = ((k + 1/2)/2^{R-1})^2.$$

Using

$$\frac{a^2 + b^2}{2} - ((a+b)/2)^2 = \frac{a^2 - 2 \cdot a \cdot b + b^2}{4} = \frac{(a-b)^2}{4}$$

we get

$$\begin{aligned} (S_{R-1}((k + 1/2)/2^{R-1}) - S_R((k + 1/2)/2^{R-1})) &= \frac{((k+1)/2^{R-1} - (k/2^{R-1})^2}{4} = \frac{1/2^{2R-2}}{4} \\ &= \frac{1}{2^{2R}}, \end{aligned}$$

which implies the assertion.
Since $S_0(x) = x$ we can recursively conclude that

$$S_R(x) = x - \sum_{s=1}^{R} \frac{g_s(x)}{2^{2s}}$$

with

$$|S_R(x) - x^2| \leq 2^{-2R-2}$$

for $x \in [0, 1]$.
In a *third step of the proof* we show, that there exists a feedforward neural network that computes $S_R(x)$ for $x \in [0, 1]$. The function $g(x)$ can be implemented by the network

$$\hat{f}_g(x) = 2 \cdot \sigma(x) - 4 \cdot \sigma(x - \frac{1}{2}) + 2 \cdot \sigma(x - 1)$$

and the function $g_s(x)$ can be implemented by a network

$$\hat{f}_{g_s} \in \mathcal{F}(s, 3)$$

with

$$\hat{f}_{g_s}(x) = \underbrace{\hat{f}_g(\hat{f}_g(\ldots(\hat{f}_g}_{s}(x))).$$

Let

$$\hat{f}_{id}(z) = \sigma(z) - \sigma(-z)$$

with

$$\begin{aligned} \hat{f}_{id}^0(z) &= z && (z \in \mathbb{R}) \\ \hat{f}_{id}^{t+1}(z) &= \hat{f}_{id}(\hat{f}_{id}^t(z)) && (z \in \mathbb{R}, t \in \mathbb{N}_0) \end{aligned}$$

be the network satisfying

$$\hat{f}_{id}^t(z) = z.$$

By combining the networks above we can implement the function $S_R(x)$ by a network

$$\hat{f}_{sq_{[0,1]}} \in \mathcal{F}(R, 7)$$

recursively defined as follows: We set $\hat{f}_{1,0}(x) = \hat{f}_{2,0}(x) = x$ and $\hat{f}_{3,0}(x) = 0$. Then we set

$$\hat{f}_{1,i+1}(x) = \hat{f}_{id}(\hat{f}_{1,i}(x)),$$

$$\hat{f}_{2,i+1}(x) = \hat{f}_g(\hat{f}_{2,i}(x))$$

and

$$\hat{f}_{3,i+1}(x) = \hat{f}_{id}(\hat{f}_{3,i}(x)) - \hat{f}_g(\hat{f}_{2,i}(x))/2^{2(i+1)}$$

for $i \in \{0, 1, \ldots, R-2\}$ and

$$\hat{f}_{sq_{[0,1]}}(x) = \hat{f}_{id}(\hat{f}_{1,R-1}(x) + \hat{f}_{3,R-1}(x)) - \hat{f}_g(\hat{f}_{2,R-1}(x))/2^{2R}.$$

By induction it is easy to see that we have

$$\hat{f}_{1,i}(x) = x, \quad \hat{f}_{2,i}(x) = g_i(x) \quad \text{and} \quad \hat{f}_{3,i}(x) = -\sum_{r=1}^{i} \frac{g_r(x)}{2^{2r}}$$

which implies

$$\hat{f}_{sq_{[0,1]}}(x) = x - \sum_{r=1}^{R-1} \frac{g_r(x)}{2^{2r}} - \frac{g_R(x)}{2^{2R}} = S_R(x),$$

hence $\hat{f}_{sq_{[0,1]}}(x)$ satisfies

$$|\hat{f}_{sq_{[0,1]}}(x) - x^2| \leq 2^{-2R-2} \tag{14}$$

for $x \in [0,1]$. It is easy to see that $\hat{f}_{sq_{[0,1]}}$ can be computed by a ReLU neural network with R layers and $2+3+2=7$ neurons per layer.
In a *last step of the proof* we show that we can also approximate the function $f(x) = x^2$ by a neural network, if $x \in [-a, a]$. Therefore let $f_{tran} : [-a, a] \to [0, 1]$ with

$$f_{tran}(z) = \frac{z}{2a} + \frac{1}{2}$$

be the function that transfers the value of $x \in [-a, a]$ in the interval, where (14) holds. Set

$$\hat{f}_{sq}(x) = 4a^2 \hat{f}_{sq_{[0,1]}}(f_{tran}(x)) - 2a \cdot \hat{f}_{id}^R(x) - a^2.$$

Since

$$x^2 = 4a^2 \cdot \left(\frac{x}{2a} + \frac{1}{2}\right)^2 - 2ax - a^2$$

we have

$$\begin{aligned} &|\hat{f}_{sq}(x) - x^2| \\ \leq &4a^2 \cdot |\hat{f}_{sq_{[0,1]}}(f_{tran}(x)) - (f_{tran}(x))^2| + 2a|\hat{f}_{id}^R(x) - x| \\ \leq &4a^2 \cdot 2^{-2R-2} = a^2 \cdot 4^{-R}. \end{aligned}$$

□

We can use the network of Lemma 2.2 to construct a network which approximates the product of two numbers.

Lemma 2.3 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the ReLU activation function $\sigma(x) = \max\{x, 0\}$. Then for any $R \in \mathbb{N}$ and any $a \geq 1$ a neural network*

$$\hat{f}_{mult} \in \mathcal{F}(R, 18)$$

exists such that

$$|\hat{f}_{mult}(x, y) - x \cdot y| \leq 2 \cdot a^2 \cdot 4^{-R}$$

holds for all $x, y \in [-a, a]$.

Proof. Let

$$\hat{f}_{sq} \in \mathcal{F}(R, 9)$$

be the neural network from Lemma 2.2 satisfying

$$|\hat{f}_{sq}(x) - x^2| \leq 4 \cdot a^2 \cdot 4^{-R}$$

for $x \in [-2a, 2a]$, and set

$$\hat{f}_{mult}(x, y) = \frac{1}{4} \cdot \left(\hat{f}_{sq}(x + y) - \hat{f}_{sq}(x - y) \right).$$

Since

$$x \cdot y = \frac{1}{4} \left((x + y)^2 - (x - y)^2 \right)$$

we have

$$\begin{aligned} |\hat{f}_{mult}(x, y) - x \cdot y| &\leq \frac{1}{4} \cdot \left| \hat{f}_{sq}(x + y) - (x + y)^2 \right| + \frac{1}{4} \cdot \left| (x - y)^2 - \hat{f}_{sq}(x - y) \right| \\ &\leq \frac{1}{4} \cdot 2 \cdot 4 \cdot a^2 \cdot 4^{-R} \\ &\leq 2 \cdot a^2 \cdot 4^{-R} \end{aligned}$$

for $x, y \in [-a, a]$. □

Next we extend the previous lemma such that the network computes the product of finitely many numbers.

Lemma 2.4 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the ReLU activation function $\sigma(x) = \max\{x, 0\}$. Then for any $a \geq 1$ and any $R \in \mathbb{N}$ with $R \geq \log_4(2 \cdot 4^{2d} \cdot a^{2d})$ a neural network*

$$\hat{f}_{mult,d} \in \mathcal{F}(R \cdot \lceil \log_2(d) \rceil, 18d)$$

exists such that

$$\left| \hat{f}_{mult,d}(\mathbf{x}) - \prod_{i=1}^{d} x^{(i)} \right| \leq 4^{4d+1} \cdot a^{4d} \cdot d \cdot 4^{-R}$$

holds for all $\mathbf{x} \in [-a, a]^d$.

Proof.
We set $q = \lceil \log_2(d) \rceil$. The feedforward neural network $\hat{f}_{mult,d}$ with $L = R \cdot q$ hidden layers and $r = 18d$ neurons in each layer is constructed as follows: Set

$$(z_1, \ldots, z_{2^q}) = \left(x^{(1)}, x^{(2)}, \ldots, x^{(d)}, \underbrace{1, \ldots, 1}_{2^q - d} \right). \tag{15}$$

In the construction of our network we will use the network $\hat{f}_{mult}$ of Lemma 2.3, which satisfies

$$|\hat{f}_{mult}(x,y) - x \cdot y| \leq 2 \cdot (4^d a^d)^2 \cdot 4^{-R} \tag{16}$$

for $x, y \in [-4^d a^d, 4^d a^d]$. In the first R layers we compute

$$\hat{f}_{mult}(z_1, z_2), \hat{f}_{mult}(z_3, z_4), \ldots, \hat{f}_{mult}(z_{2^q-1}, z_{2^q}),$$

which can be done by R layers of $18 \cdot 2^{q-1} \leq 18 \cdot d$ neurons. E.g., in case in case $z_l = x^{(d)}$ and $z_{l+1} = 1$ we have

$$\hat{f}_{mult}(z_l, z_{l+1}) = \hat{f}_{mult}(x^{(d)}, 1).$$

As a result of the first R layers we get a vector of outputs which has length 2^{q-1}. Next we pair these outputs and apply $\hat{f}_{mult}$ again. This procedure is continued until there is only one output left. Therefore we need $L = Rq$ hidden layers and at most $18d$ neurons in each layer.

By (16) and $R \geq \log_4 \left(2 \cdot 4^{2 \cdot d} \cdot a^{2 \cdot d}\right)$ we get for any $l \in \{1, \ldots, d\}$ and any $z_1, z_2 \in [-(4^l - 1) \cdot a^l, (4^l - 1) \cdot a^l]$

$$|\hat{f}_{mult}(z_1, z_2)| \leq |z_1 \cdot z_2| + |\hat{f}_{mult}(z_1, z_2) - z_1 \cdot z_2| \leq (4^l - 1)^2 a^{2l} + 1 \leq (4^{2l} - 1) \cdot a^{2l}.$$

From this we get successively that all outputs of layer $l \in \{1, \ldots, q-1\}$ are contained in the interval $[-(4^{2^l} - 1) \cdot a^{2^l}, (4^{2^l} - 1) \cdot a^{2^l}]$, hence in particular they are contained in the interval $[-4^d a^d, 4^d a^d]$ where inequality (16) does hold.

Define $\hat{f}_{2^q}$ recursively by

$$\hat{f}_{2^q}(z_1, \ldots, z_{2^q}) = \hat{f}_{mult}(\hat{f}_{2^{q-1}}(z_1, \ldots, z_{2^{q-1}}), \hat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \ldots, z_{2^q}))$$

and

$$\hat{f}_2(z_1, z_2) = \hat{f}_{mult}(z_1, z_2),$$

and set

$$\Delta_l = \sup_{z_1, \ldots, z_{2^l} \in [-a,a]} |\hat{f}_{2^l}(z_1, \ldots, z_{2^l}) - \prod_{i=1}^{2^l} z_i|.$$

Then

$$|\hat{f}_{mult,d}(\mathbf{x}) - \prod_{i=1}^{d} x^{(i)}| \leq \Delta_q$$

and from

$$\Delta_1 \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R}$$

(which follows from (16)) and

$$\begin{aligned}
\Delta_q \;\leq\; & \sup_{z_1,\ldots,z_{2^q}\in[-a,a]} |\hat{f}_{mult}(\hat{f}_{2^{q-1}}(z_1,\ldots,z_{2^{q-1}}),\hat{f}_{2^{q-1}}(z_{2^{q-1}+1},\ldots,z_{2^q})) \\
& \qquad -\hat{f}_{2^{q-1}}(z_1,\ldots,z_{2^{q-1}})\cdot\hat{f}_{2^{q-1}}(z_{2^{q-1}+1},\ldots,z_{2^q})| \\
& + \sup_{z_1,\ldots,z_{2^q}\in[-a,a]} \Big| \hat{f}_{2^{q-1}}(z_1,\ldots,z_{2^{q-1}})\cdot\hat{f}_{2^{q-1}}(z_{2^{q-1}+1},\ldots,z_{2^q}) \\
& \qquad -\left(\prod_{i=1}^{2^{q-1}} z_i\right)\cdot\hat{f}_{2^{q-1}}(z_{2^{q-1}+1},\ldots,z_{2^q})\Big| \\
& + \sup_{z_1,\ldots,z_{2^q}\in[-a,a]} \Big| \left(\prod_{i=1}^{2^{q-1}} z_i\right)\cdot\hat{f}_{2^{q-1}}(z_{2^{q-1}+1},\ldots,z_{2^q}) \\
& \qquad -\left(\prod_{i=1}^{2^{q-1}} z_i\right)\cdot\prod_{i=2^{q-1}+1}^{2^q} z_i\Big| \\
\leq\; & 2\cdot(4^d\cdot a^d)^2\cdot 4^{-R} + 2\cdot 4^{2^{q-1}}\cdot a^{2^{q-1}}\cdot\Delta_{q-1}
\end{aligned}$$

(where the last inequality follows from (16) and the fact that all outputs of layer $l\in\{1,\ldots,q-1\}$ are contained in the interval $[-4^{2^l}a^{2^l},4^{2^l}a^{2^l}]$) we get for $x\in[-a,a]^d$

$$\begin{aligned}
& |\hat{f}_{mult,d}(\mathbf{x})-\prod_{i=1}^{d} x^{(i)}| \\
& \leq \Delta_q \\
& \leq 2\cdot(4^d\cdot a^d)^2\cdot 4^{-R}\cdot 4^{1+2+\cdots+2^{q-1}}\cdot a^{1+2+\cdots+2^{q-1}}\cdot\left(1+2+\cdots+2^{q-1}\right) \\
& \leq (4^d\cdot a^d)^2\cdot 4^{-R}\cdot 4^{2d+1}\cdot a^{2d}\cdot d \\
& = 4^{4d+1}\cdot a^{4d}\cdot d\cdot 4^{-R},
\end{aligned}$$

where the last inequality was implied by

$$1+2+\cdots+2^{q-1}=2^q\leq 2\cdot d.$$

□

We can now formulate and prove our main result.

Theorem 2.1 *Let $p,C>0$, $A\geq 1$ and $K\in\mathbb{N}$ be arbitrary and define the space $\mathcal{G}$ of neural networks by*

$$\mathcal{G}=\left\{\sum_{k=1}^{(K+1)^d}\sum_{\substack{j_1,\ldots,j_d\in\{0,\ldots,\lceil p\rceil\}\\ j_1+\cdots+j_d\leq p}} f_{k,j_1,\ldots,j_d} \;:\; f_{k,j_1,\ldots,j_d}\in\mathcal{F}(L,r)\right\}$$

where $\mathcal{F}(L,r)$ is the space of all neural networks with L hidden layers, r neurons per layer and ReLU activation function, and where

$$L=\max\left\{\lceil(p+d)\cdot\log_4 K\rceil,\lceil\log_4(2\cdot(8\cdot A)^{2p+2d})\rceil\right\}\cdot\lceil\log_2(p+d)\rceil+1$$

and

$$r = 18 \cdot \lceil p \rceil + 18 \cdot d.$$

Then for any (p, C)–smooth function $g : \mathbb{R}^d \to \mathbb{R}$ it holds

$$d(g, \mathcal{G}, \|\cdot\|_{\infty,[-A,A]^d}) \leq c_1 \cdot \frac{1}{K^p}.$$

Proof. Let $p = q + \beta$ for some $q \in \mathbb{N}_0$ and $\beta \in (0, 1]$, and let

$$\begin{aligned} P(x) \;=\; & \sum_{k=1}^{(K+1)^d} \sum_{\substack{j_1,\ldots,j_d \in \{0,\ldots,q\} \\ j_1+\cdots+j_d \leq q}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} g}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_{\mathbf{i}_k}) \cdot \prod_{i=1}^{d} (x^{(i)} - x^{(i)}_{\mathbf{i}_k})^{j_i} \\ & \cdot \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}|\right)_+ \end{aligned}$$

be the local convex combination of Taylor polynomials of g introduced in (13). It suffices to show that there exists $f \in \mathcal{G}$ such that

$$|P(x) - f(x)| \leq c_1 \cdot \frac{1}{K^p}$$

holds for any $x \in [-A, A]^d$.
To show this, it suffices to show that for any $k \in \{1, \ldots, (K+1)^d\}$ and $j_1, \ldots, j_d \in \{0, \ldots, q\}$ with $j_1 + \cdots + j_d \leq q$ there exists $f_{k,j_1,\ldots,j_d} \in \mathcal{F}(L, r)$ such that

$$\begin{aligned} \Bigg| \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} g}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_{\mathbf{i}_k}) \cdot \prod_{i=1}^{d} (x^{(i)} - x^{(i)}_{\mathbf{i}_k})^{j_i} \cdot \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}|\right)_+ \\ - f_{k,j_1,\ldots,j_d}(x) \Bigg| \leq \frac{c_2}{K^{p+d}} \end{aligned}$$

holds for all $x \in [-A, A]^d$.
Because of

$$\begin{aligned} & \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}|\right)_+ \\ & = \max\left\{0, \frac{K}{2A} \cdot \left(x - \left(x^{(j)}_{\mathbf{i}_k} - \frac{2A}{K}\right)\right)\right\} - 2 \cdot \max\left\{0, \frac{K}{2A} \cdot \left(x - x^{(j)}_{\mathbf{i}_k}\right)\right\} \\ & \quad + \max\left\{0, \frac{K}{2A} \cdot \left(x - \left(x^{(j)}_{\mathbf{i}_k} + \frac{2A}{K}\right)\right)\right\} \end{aligned}$$

we can compute with one layer of $2 \cdot q + 3 \cdot d$ many neurons the values

$$(x^{(i)} - x^{(i)}_{\mathbf{i}_k}) = \max\left\{0, x^{(i)} - x^{(i)}_{\mathbf{i}_k}\right\} + \max\left\{0, x^{(i)}_{\mathbf{i}_k} - x^{(i)}\right\}$$

j_i-times for each $j = 1, \ldots, d$ and the values

$$\left(1 - \frac{K}{2A} \cdot |x^{(j)} - x^{(j)}_{\mathbf{i}_k}|\right)_+$$

for each $j = 1, \ldots, d$. Then we apply Lemma 2.4 (with an obvious modification of the output layer and with sufficiently large $R \geq \log_4(K^{p+d})$) to compute

$$\frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} g}{\partial^{j_1} x^{(1)} \cdots \partial^{j_d} x^{(d)}}(x_{\mathbf{i}_k}) \cdot \prod_{i=1}^{d} (x^{(i)} - x_{\mathbf{i}_k}^{(i)})^{j_i} \cdot \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x_{\mathbf{i}_k}^{(j)}|\right)_+ .$$

Here the modification of the output layer ensures that we only have to compute

$$\prod_{i=1}^{d} (x^{(i)} - x_{\mathbf{i}_k}^{(i)})^{j_i} \cdot \prod_{j=1}^{d} \left(1 - \frac{K}{2A} \cdot |x^{(j)} - x_{\mathbf{i}_k}^{(j)}|\right)_+ ,$$

and we represent

$$\prod_{i=1}^{d} (x^{(i)} - x_{\mathbf{i}_k}^{(i)})^{j_i}$$

by a product of at most q terms of the form $x^{(i)} - x_{\mathbf{i}_k}^{(i)}$. These terms take on values in $[-2A, 2A]$, all the remaining ones are contained in $[0, 1] \subseteq [-2A, 2A]$, so we have a product of $q + d$ factors contained in $[-2A, 2A]$. So we can apply Lemma 2.4 with $d = q + d$ and $a = 2 \cdot A$. Here the assumption $R \geq \log_4(2 \cdot 4^{2d} \cdot a^{2d})$ means that we need $R \geq \log_4(2 \cdot (8A)^{2d+2q})$. The network has then not R but $R \cdot \lceil \log_2(d + q) \rceil$ many layers and $18 \cdot (q + d)$ neurons per layer, and because of our first layer we need to add one more layer. Application of Lemma 2.4 yields the assertion. □

Remark. The neural network in the space $\mathcal{G}$ in Theorem 2.1 have

$$c_2 \cdot K^d \cdot \log K$$

many weights and are able to approximate a (p, C)–smooth function with an error of order

$$\frac{1}{K^p}.$$

So up to a logarithmic factor this is the same result as we have shown in Subsection 2.1 for piecewise polynomials. But, as we will see later, the advantage of our result for neural networks is that *due to the network structure we can derive from this result nice results concerning the approximation of compositions of functions.*

3 Neural network generalization

In this chapter we want to bound the *generalization error* of neural network estimates, which is this part of the error which arises because the neural network estimate is adopted to some *empirical risk* defined by a sample average instead of the *risk* defined by an expectation. To do this we will use results from the so–called *VC theory* (Vapnik-Chervonenkis theory).

3.1 Motivation

Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\{Y^2\} < \infty$. Let

$$m : \mathbb{R}^d \to \mathbb{R}, \quad m(x) = \mathbf{E}\{Y|X = x\}$$

be the corresponding regression function. Because of

$$\begin{aligned} \mathbf{E}\{|Y - f(X)|^2\} &= \mathbf{E}\{((Y - m(X)) + (m(X) - f(X)))^2\} \\ &= \mathbf{E}\{|Y - m(X)|^2\} + \int |f(x) - m(x)|^2 \mathbf{P}_X(dx), \end{aligned}$$

where the last equality follows in case $\mathbf{E}\{f(X)^2\} < \infty$ from

$$\begin{aligned} &\mathbf{E}\{((Y - m(X)) \cdot (m(X) - f(X)))\} \\ &= \mathbf{E}\{\mathbf{E}\{\cdot|X\}\} = \mathbf{E}\{(m(X) - f(X)) \cdot \mathbf{E}\{(Y - m(X))|X\}\} \\ &= \mathbf{E}\{(m(X) - f(X)) \cdot (\mathbf{E}\{Y|X\} - m(X))\} = 0 \end{aligned}$$

(and trivially holds in case $\mathbf{E}\{f(X)^2\} = \infty$, since then both sides are equal to ∞) we have

$$m(\cdot) = \arg \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}\{|f(X) - Y|^2\}.$$

Let $\mathcal{F}_n$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$, set

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

and let

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2$$

be the corresponding *least squares estimate* of the regression function m based on the sample $\mathcal{D}_n$ of (X, Y).
The aim in the sequel is to bound the so-called L_2 *error*

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) = \mathbf{E}\left\{|m_n(X) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\}$$

of the least squares estimate.

It is easy to bound an empirical version of it, because by definition of the estimate we have:

$$\begin{aligned} Z_n &:= \frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2 \\ &= \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2, \end{aligned}$$

which implies

$$\begin{aligned} \mathbf{E}\{Z_n\} &\leq \min_{f\in\mathcal{F}_n}\mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\ &= \min_{f\in\mathcal{F}_n}\int |f(x)-m(x)|^2\mathbf{P}_X(dx). \end{aligned}$$

In the sequel we will derive upper bounds on the difference between the L_2 error and (some factor times) its empirical version.

3.2 Uniform exponential inequalities

In Section 3.1 we need upper bounds for terms like

$$\mathbf{E}\left\{|m_n(X)-Y|^2\big|\mathcal{D}_n\right\} - \frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2.$$

One problem here is, that within the expectation and the sample average there occurs a *random* function $m_n \in \mathcal{F}_n$. To get rid of this problem, one can upper bound the above term by

$$\sup_{f\in\mathcal{F}_n}\left\{\mathbf{E}\left\{|f(X)-Y|^2\right\} - \frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2\right\}.$$

In order to derive upper bounds for terms like this, we need a measure for the "complexity" of the function space $\mathcal{F}_n$, which we introduce next.

Definition 3.1 *Let $\epsilon > 0$, let $\mathcal{G}$ a set of functions $g : \mathbb{R}^l \to \mathbb{R}$, let $1 \leq p < \infty$ and let ν be a probability measure on $\mathbb{R}^l$. For $g : \mathbb{R}^l \to \mathbb{R}$ set*

$$\|g\|_{L_p(\nu)} := \left\{\int |g(x)|^p \nu(dx)\right\}^{\frac{1}{p}}.$$

a) *A finite set of functions $g_1, \ldots, g_N : \mathbb{R}^l \to \mathbb{R}$ satisfying*

$$\forall g \in \mathcal{G}\ \exists j = j(g) \in \{1,\ldots,N\} :\ \|g - g_j\|_{L_p(\nu)} < \epsilon$$

is called **ϵ-cover of $\mathcal{G}$** *w.r.t. $\|\cdot\|_{L_p(\nu)}$.*
b) *The* **ϵ-covering number of $\mathcal{G}$** *w.r.t. $\|\cdot\|_{L_p(\nu)}$*

$$\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)})$$

is defined as the size of the smallest ϵ-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$. In case that there does not exist a finite ϵ-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$ we set $\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}) = \infty$.
c) *Let $z_1^n = (z_1, \ldots, z_n)$ be n points in $\mathbb{R}^l$. Let ν_n be the corresponding empirical distribution, i.e.,*

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(z_i) \quad (A \subseteq \mathbb{R}^l),$$

which implies

$$\|g\|_{L_p(\nu_n)} = \left\{\frac{1}{n}\sum_{i=1}^{n}|g(z_i)|^p\right\}^{\frac{1}{p}}.$$

Any ϵ-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu_n)}$ is called **L_p-ϵ-cover $\mathcal{G}$ on z_1^n**, *and for the ϵ-covering number of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu_n)}$ we use the notation*

$$\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n).$$

$\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n)$ *is called* **L_p-ϵ-covering number of $\mathcal{G}$ on z_1^n**.

Theorem 3.1 *(Pollard (1984)).*
Let Z, Z_1, …, Z_n be i.i.d. $\mathbb{R}^l$-valued random variables. Let $B > 0$ and let $\mathcal{G}$ be a class of functions $g : \mathbb{R}^l \to [0, B]$. Then it holds for every $n \in \mathbb{N}$ and every $\epsilon > 0$:

$$\begin{aligned}&\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbf{E}\{g(Z)\}\right| > \epsilon\right\}\\ &\leq 8 \cdot \mathbf{E}\left\{\mathcal{N}_1(\epsilon/8, \mathcal{G}, Z_1^n)\right\} \cdot \exp\left(-\frac{n\cdot\epsilon^2}{128\cdot B^2}\right),\end{aligned}$$

where $Z_1^n = (Z_1, \ldots, Z_n)$.

Remark: In Theorem 3.1 we ignore possible measurability problems (which can occur in connection with the supremum or in connection with the covering number).

Proof of Theorem 3.1. The proof will be divided into four steps.
Step 1: Symmetrization by a ghost sample
We will replace the expectation inside the above probability by an empirical mean of a "ghost sample". To do this, we let Z_1', …, Z_n' be i.i.d. random variables distributed as Z_1 and independent of Z_1^n and set

$$Z'^n_1 = (Z_1' \ldots, Z_n').$$

Let $g^* = g^*(Z_1^n)$ be a function $g \in \mathcal{G}$ such that

$$\left|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbf{E}\{g(Z)\}\right| > \epsilon$$

if there exists such a function, and let g^* be an arbitrary function contained in $\mathcal{G}$ otherwise.

Because of $0 \leq g^*(Z) \leq B$ we have

$$\begin{aligned}\mathbf{V}\{g^*(Z)|Z_1^n\} &= \mathbf{V}\left\{g^*(Z) - \frac{B}{2}\big|Z_1^n\right\} \\ &\leq \mathbf{E}\left\{\left|g^*(Z) - \frac{B}{2}\right|^2 \big|Z_1^n\right\} \leq \frac{B^2}{4}.\end{aligned}$$

From this we can conclude by Chebyshev's inequality

$$\begin{aligned}&\mathbf{P}\left\{\mathbf{E}\{g^*(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g^*(Z_i') > \frac{\epsilon}{2}|Z_1^n\right\} \\ &\leq \frac{\mathbf{V}\{g^*(Z)|Z_1^n\}}{n \cdot \left(\frac{\epsilon}{2}\right)^2} \leq \frac{\frac{B^2}{4}}{n \cdot \frac{\epsilon^2}{4}} = \frac{B^2}{n \cdot \epsilon^2}.\end{aligned}$$

Thus, for $n \geq \frac{2 \cdot B^2}{\epsilon^2}$, we have

$$\mathbf{P}\left\{\mathbf{E}\{g^*(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g^*(Z_i') \leq \frac{\epsilon}{2}|Z_1^n\right\} \geq \frac{1}{2},$$

which implies

$$\begin{aligned}&\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i')\right| > \frac{\epsilon}{2}\right\} \\ &\geq \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \frac{1}{n}\sum_{i=1}^n g^*(Z_i')\right| > \frac{\epsilon}{2}\right\} \\ &\geq \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \mathbf{E}\{g^*(Z)|Z_1^n\}\right| > \epsilon, \left|\frac{1}{n}\sum_{i=1}^n g^*(Z_i') - \mathbf{E}\{g^*(Z)|Z_1^n\}\right| \leq \frac{\epsilon}{2}\right\} \\ &= \mathbf{E}\left\{I_{\left\{\left|\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \mathbf{E}\{g^*(Z)|Z_1^n\}\right| > \epsilon\right\}} \cdot \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n g^*(Z_i') - \mathbf{E}\{g^*(Z)|Z_1^n\}\right| \leq \frac{\epsilon}{2}|Z_1^n\right\}\right\} \\ &\geq \frac{1}{2} \cdot \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \mathbf{E}\{g^*(Z)|Z_1^n\}\right| > \epsilon\right\} \\ &= \frac{1}{2} \cdot \mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbf{E}\{g(Z)\}\right| > \epsilon\right\}.\end{aligned}$$

This proves

$$\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbf{E}\{g(Z)\}\right| > \epsilon\right\} \leq 2 \cdot \mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i')\right| > \frac{\epsilon}{2}\right\}$$

for $n \geq \frac{2 \cdot B^2}{\epsilon^2}$.

Step 2: Introduction of additional randomness by random signs.

Let $U_1, \ldots, U_n$ be independent and uniformly distributed on $\{-1, 1\}$ and independent of $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$. Because of $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$ i.i.d. the joint distribution of

Z_1^n, Z'^n_1 is not affected if one randomly interchanges (corresponding) components of Z_1^n and Z'^n_1. Hence

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)-\frac{1}{n}\sum_{i=1}^n g(Z_i')\right|>\frac{\epsilon}{2}\right\}\\
&=\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot(g(Z_i)-g(Z_i'))\right|>\frac{\epsilon}{2}\right\}\\
&\le\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(Z_i)\right|>\frac{\epsilon}{4}\right\}+\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(Z_i')\right|>\frac{\epsilon}{4}\right\}\\
&=2\cdot\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(Z_i)\right|>\frac{\epsilon}{4}\right\}.
\end{aligned}$$

Step 3: Conditioning and introduction of a covering.
Because of U_1^n and Z_1^n independent we have

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(Z_i)\right|>\frac{\epsilon}{4}\right\}\\
&=\mathbf{E}\left\{\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(Z_i)\right|>\frac{\epsilon}{4}\middle|Z_1^n\right\}\right\}\\
&\le\int_{(\mathbb{R}^d)^n}\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(z_i)\right|>\frac{\epsilon}{4}\right\}\mathbf{P}_{Z_1^n}(dz_1^n).
\end{aligned}$$

Fix $z_1,\ldots,z_n\in\mathbb{R}^d$ and let $\mathcal{G}_{\epsilon/8}$ be an L_1-$\frac{\epsilon}{8}$-cover of $\mathcal{G}$ on z_1^n of minimal size

$$|\mathcal{G}_{\epsilon/8}|=\mathcal{N}_1(\epsilon/8,\mathcal{G},z_1^n).$$

W.l.o.g. we may assume $0\le\bar{g}(z)\le B$ for all $\bar{g}\in\mathcal{G}_{\epsilon/8}$ (otherwise, truncate $\bar{g}$ correspondingly).
For any $g\in\mathcal{G}$ there exists $\bar{g}\in\mathcal{G}_{\epsilon/8}$ such that

$$\frac{1}{n}\sum_{i=1}^n|g(z_i)-\bar{g}(z_i)|<\frac{\epsilon}{8},$$

which implies

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n U_i\cdot g(z_i)\right| &= \left|\frac{1}{n}\sum_{i=1}^n U_i\cdot\bar{g}(z_i)+\frac{1}{n}\sum_{i=1}^n U_i\cdot(g(z_i)-\bar{g}(z_i))\right|\\
&\le \left|\frac{1}{n}\sum_{i=1}^n U_i\cdot\bar{g}(z_i)\right|+\frac{1}{n}\sum_{i=1}^n|g(z_i)-\bar{g}(z_i)|\\
&\le \left|\frac{1}{n}\sum_{i=1}^n U_i\cdot\bar{g}(z_i)\right|+\frac{\epsilon}{8}.
\end{aligned}$$

Using this bound and the union bound we can conclude

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|>\frac{\epsilon}{4}\right\}\\
&\leq \mathbf{P}\left\{\sup_{g\in\mathcal{G}_{\epsilon/8}}\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|+\frac{\epsilon}{8}>\frac{\epsilon}{4}\right\}\\
&= \mathbf{P}\left\{\exists g\in\mathcal{G}_{\epsilon/8}:\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|>\frac{\epsilon}{8}\right\}\\
&\leq |\mathcal{G}_{\epsilon/8}|\cdot\max_{g\in\mathcal{G}_{\epsilon/8}}\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|>\frac{\epsilon}{8}\right\}\\
&= \mathcal{N}_1(\epsilon/8,\mathcal{G},z_1^n)\cdot\max_{g\in\mathcal{G}_{\epsilon/8}}\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|>\frac{\epsilon}{8}\right\}.
\end{aligned}$$

Step 4: Application of Hoeffding's inequality
In this step we bound

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|>\frac{\epsilon}{8}\right\}$$

where $z_1,\dots,z_n\in\mathbb{R}^d$, $g:\mathbb{R}^d\to\mathbb{R}$ and $0\leq g(z)\leq B$.
Since

$$U_1\cdot g(z_1),\dots,U_n\cdot g(z_n)$$

are independent random variables with expectation zero and

$$-B\leq U_i\cdot g(z_i)\leq B\quad(i=1,\dots,n)$$

we have by Hoeffding's inequality

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i\cdot g(z_i)\right|>\frac{\epsilon}{8}\right\}\leq 2\cdot\exp\left(-\frac{2\cdot n\cdot\left(\frac{\epsilon}{8}\right)^2}{(2\cdot B)^2}\right)=2\cdot\exp\left(-\frac{n\cdot\epsilon^2}{128\cdot B^2}\right).$$

In case $n\geq 2\cdot B^2/\epsilon^2$ the assertion is now implied by the above four steps. For $n<2\cdot B^2/\epsilon^2$ the bound on the probability trivially holds, since the right-hand side is greater than one.

□

The right-hand side in the upper bound on the probability in Theorem 3.1 does not converge to zero in case $\epsilon\leq 1/\sqrt{n}$, which is in view of the optimal rate of convergence of regression estimates not sufficient. In order to derive upper bounds which converge faster against zero, one can consider the difference between expectations and a factor greater

than one times the corresponding sample average, because we have

$$\begin{aligned}
& \mathbf{E}\left\{|m_n(X)-Y|^2\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
& \qquad -2\cdot\frac{1}{n}\sum_{i=1}^{n}\left(|m_n(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right) \\
& > t \\
\Leftrightarrow \quad & \mathbf{E}\left\{|m_n(X)-Y|^2\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
& \qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|m_n(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right) \\
& > \frac{1}{2}\cdot\left(t + \mathbf{E}\left\{|m_n(X)-Y|^2\big|\mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\}\right)
\end{aligned}$$

and the following theorem (which we can apply with $\epsilon = 1/2$ and $\alpha = \beta = t/2$).

Theorem 3.2 *(Lee, Bartlett and Williamson (1996)).*
Let (X,Y), (X_1,Y_1), $\ldots$, (X_n,Y_n) be i.i.d. $\mathbb{R}^d \times \mathbb{R}$-valued random variables with $|Y| \le B$ a.s. for some $B \ge 1$. Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [-B,B]$. Then we have for any $n \in \mathbb{N}$, $\alpha, \beta > 0$ and any $0 < \epsilon \le 1/2$:

$$\begin{aligned}
& \mathbf{P}\Bigg\{\exists f \in \mathcal{F} : \mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
& \qquad\qquad -\frac{1}{n}\sum_{i=1}^{n}\left(|f(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right) \\
& \qquad > \epsilon\cdot\left(\alpha+\beta + \mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg\} \\
& \le 14\cdot\sup_{x_1^n}\mathcal{N}_1\left(\frac{\beta\cdot\epsilon}{20\cdot B}, \mathcal{F}, x_1^n\right)\cdot\exp\left(-\frac{\epsilon^2(1-\epsilon)\cdot\alpha\cdot n}{214\cdot(1+\epsilon)\cdot B^4}\right).
\end{aligned}$$

In the proof of Theorem 3.2 we will need the following auxiliary result, which will be proven in the practising course.

Theorem 3.3 *Let $B \ge 1$ and let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to [0,B]$. Let Z, Z_1, $\ldots$, Z_n be i.i.d. $\mathbb{R}^d$–valued random variables. Assume $\alpha > 0$, $0 < \epsilon > 1$, and $n \ge 1$. Then*

$$\begin{aligned}
& \mathbf{P}\left\{\sup_{g\in\mathcal{G}}\frac{\frac{1}{n}\cdot\sum_{i=1}^{n}g(Z_i) - \mathbf{E}g(Z)}{\alpha + \frac{1}{n}\cdot\sum_{i=1}^{n}g(Z_i) + \mathbf{E}g(Z)} > \epsilon\right\} \\
& \le 4\cdot\mathbf{E}\mathcal{N}_1\left(\frac{\alpha\cdot\epsilon}{5}, \mathcal{G}, Z_1^n\right)\cdot\exp\left(-\frac{3\cdot\epsilon^2\cdot\alpha\cdot n}{40\cdot B}\right).
\end{aligned}$$

Proof of Theorem 3.2: Let us introduce the following notation

$$Z = (X,Y),\ Z_i = (X_i,Y_i),\ i = 1,\ldots,n,$$

and

$$g_f(x,y) = |f(x) - y|^2 - |m(x) - y|^2.$$

Observe that $|f(x)| \le B$, $|y| \le B$ and $|m(x)| \le B$ imply

$$-4B^2 \le g_f(x,y) \le 4B^2.$$

We can rewrite the probability in the theorem as follows

$$\mathbf{P}\left\{\exists f \in \mathcal{F} : \mathbf{E}g_f(Z) - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) \ge \epsilon(\alpha + \beta + \mathbf{E}g_f(Z))\right\}. \tag{17}$$

The proof will proceed in several steps.

Step 1. Symmetrization by a ghost sample.

Replace the expectation on the left side of inequality in (17) by the empirical mean based on the ghost sample Z'^n_1 of i.i.d. random variables distributed as Z and independent of Z_1^n. Consider a function $f_n \in \mathcal{F}$ depending upon Z_1^n such that

$$\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i) \ge \epsilon(\alpha + \beta) + \epsilon\mathbf{E}\{g_{f_n}(Z)|Z_1^n\},$$

if such a function exists in $\mathcal{F}$, otherwise choose an arbitrary function in $\mathcal{F}$. Chebychev's inequality together with

$$\begin{aligned}
\mathbf{V}\left\{g_{f_n}(Z)|Z_1^n\right\} &\le \mathbf{E}\left\{g_{f_n}(Z)^2|Z_1^n\right\} \\
&= \mathbf{E}\left\{\left((f_n(X) - m(X)) \cdot (f_n(X) - Y + m(X) - Y)\right)^2 |Z_1^n\right\} \\
&\le 16B^2 \cdot \mathbf{E}\left\{(f_n(X) - m(X))^2|Z_1^n\right\} \\
&= 16B^2 \cdot \mathbf{E}\left\{|f_n(X) - Y|^2 - |m(X) - Y|^2|Z_1^n\right\} \\
&= 16B^2 \cdot \mathbf{E}\left\{g_{f_n}(Z)|Z_1^n\right\}
\end{aligned}$$

imply

$$\begin{aligned}
&\mathbf{P}\Bigg\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i') \\
&\qquad\qquad > \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\Bigg| Z_1^n\Bigg\} \\
&\le \frac{\mathbf{V}\left\{g_{f_n}(Z)|Z_1^n\right\}}{n \cdot \left(\frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbf{E}\left\{g_{f_n}(Z)|Z_1^n\right\}\right)^2} \\
&\le \frac{16B^2\mathbf{E}\left\{g_{f_n}(Z)|Z_1^n\right\}}{n \cdot \left(\frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbf{E}\left\{g_{f_n}(Z)|Z_1^n\right\}\right)^2} \\
&\le \frac{16B^2}{n \cdot \left(\frac{\epsilon}{2}\right)^2} \cdot \frac{\mathbf{E}\left\{g_{f_n}(Z)|Z_1^n\right\}}{\left((\alpha + \beta) + \frac{\epsilon}{2}\mathbf{E}\left\{g_{f_n}(Z)|Z_1^n\right\}\right)^2} \\
&\le \frac{16B^2}{\epsilon^2(\alpha + \beta)n},
\end{aligned}$$

where the last inequality follows from

$$f(x) = \frac{x}{(a+x)^2} \leq f(a) = \frac{1}{4a}$$

for all $x \geq 0$ and all $a > 0$ (which holds since the derivative of f changes at $x = a$ its sign from plus to minus). Thus for $n > \frac{128B^2}{\epsilon^2(\alpha+\beta)}$

$$\mathbf{P}\left\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i') \leq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\middle| Z_1^n\right\}$$

$$\geq \frac{7}{8}. \tag{18}$$

Hence

$$\begin{aligned}
&\mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}g_f(Z)\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i') - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \geq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\right\} \\
&\geq \mathbf{P}\left\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \geq \epsilon(\alpha+\beta) + \epsilon\mathbf{E}\{g_{f_n}(Z)|Z_1^n\},\right. \\
&\qquad \left.\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_f(Z_i') \leq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\right\} \\
&= \mathbf{E}\left\{I_{\left\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \geq \epsilon(\alpha+\beta) + \epsilon\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\right\}}\right. \\
&\qquad \left.\cdot\mathbf{E}\left\{I_{\left\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_f(Z_i') \leq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\right\}}|Z_1^n\right\}\right\} \\
&= \mathbf{E}\left\{I_{\{\cdots\}}\mathbf{P}\left\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_f(Z_i')\right.\right. \\
&\qquad\qquad \left.\left.\leq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\middle| Z_1^n\right\}\right\} \\
&\geq \frac{7}{8}\mathbf{P}\left\{\mathbf{E}\{g_{f_n}(Z)|Z_1^n\} - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \geq \epsilon(\alpha+\beta) + \epsilon\mathbf{E}\{g_{f_n}(Z)|Z_1^n\}\right\} \\
&= \frac{7}{8}\mathbf{P}\left\{\exists f \in \mathcal{F} : \mathbf{E}g_f(Z) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \epsilon(\alpha+\beta) + \epsilon\mathbf{E}g_f(Z)\right\}
\end{aligned}$$

where the last inequality follows from (18). Thus we have shown that for $n > \frac{128B^2}{\epsilon^2(\alpha+\beta)}$

$$\begin{aligned}
&\mathbf{P}\left\{\exists f \in \mathcal{F} : \mathbf{E}g_f(Z) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \epsilon(\alpha+\beta) + \epsilon\mathbf{E}g_f(Z)\right\} \\
&\leq \frac{8}{7}\mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i)\right. \\
&\qquad\qquad \left.\geq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}g_f(Z)\right\}.
\end{aligned} \tag{19}$$

STEP 2. Replacement of the expectation

$$\mathbf{E}g_f(Z) \geq \frac{1}{16B^2} \cdot \mathbf{E}\{g_f(Z)^2\}$$

in (19) by an empirical mean of the ghost sample.

First we introduce additional conditions in the probability (19)

$$\begin{aligned}
&\mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}g_f(Z)\right\} \\
&\leq \mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbf{E}g_f(Z),\right. \\
&\qquad \frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbf{E}g_f^2(Z) \leq \epsilon\left(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) + \mathbf{E}g_f^2(Z)\right), \\
&\qquad \left.\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') - \mathbf{E}g_f^2(Z) \leq \epsilon\left(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') + \mathbf{E}g_f^2(Z)\right)\right\} \\
&\qquad +2\mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbf{E}g_f^2(Z)}{\left(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) + \mathbf{E}g_f^2(Z)\right)} > \epsilon\right\}. \qquad (20)
\end{aligned}$$

Application of Theorem 3.3 to the second probability on the right–hand side of (20) yields

$$\begin{aligned}
&\mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbf{E}g_f^2(Z)}{\left(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) + \mathbf{E}g_f^2(Z)\right)} > \epsilon\right\} \\
&\leq 4\mathbf{E}\mathcal{N}_1\left(\frac{(\alpha+\beta)\epsilon}{5}, \{g_f : f \in \mathcal{F}\}, Z_1^n\right)\exp\left(-\frac{3\epsilon^2(\alpha+\beta)n}{40(16B^4)}.\right)
\end{aligned}$$

Now we consider the first probability on the right side of (20). The second inequality inside the probability implies

$$(1+\epsilon)\mathbf{E}g_f^2(Z) \geq (1-\epsilon)\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \epsilon(\alpha+\beta),$$

which is equivalent to

$$\frac{1}{32B^2}\mathbf{E}g_f^2(Z) \geq \frac{1-\epsilon}{32B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \epsilon\frac{(\alpha+\beta)}{32B^2(1+\epsilon)}.$$

We can deal similarly with the third inequality. Using this and the inequality $\mathbf{E}g_f(Z) \geq \frac{1}{16B^2}\mathbf{E}g_f^2(Z) = 2\frac{1}{32B^2}\mathbf{E}g_f^2(Z)$ we can bound the first probability on the right side of (20) by

$$\begin{aligned}
&\mathbf{P}\left\{\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i)\right. \\
&\qquad \geq \epsilon(\alpha+\beta)/2 + \frac{\epsilon}{2}\left(\frac{1-\epsilon}{32B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \frac{\epsilon(\alpha+\beta)}{32B^2(1+\epsilon)}\right. \\
&\qquad\qquad \left.\left.+\frac{1-\epsilon}{32B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') - \frac{\epsilon(\alpha+\beta)}{32B^2(1+\epsilon)}\right)\right\}.
\end{aligned}$$

This shows

$$\mathbf{P}\left\{\exists f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^{n}g_f(Z_i')-\frac{1}{n}\sum_{i=1}^{n}g_f(Z_i)\geq\epsilon(\alpha+\beta)/2+\epsilon\mathbf{E}g_f(Z)/2\right\}$$
$$\leq\mathbf{P}\left\{\exists f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^{n}(g_f(Z_i')-g_f(Z_i))\right.$$
$$\left.\geq\epsilon(\alpha+\beta)/2-\frac{\epsilon^2(\alpha+\beta)}{32B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}(g_f^2(Z_i)+g_f^2(Z_i'))\right\}$$
$$+8\mathbf{E}\mathcal{N}_1\left(\frac{(\alpha+\beta)\epsilon}{5},\{g_f:f\in\mathcal{F}\},Z_1^n\right)\exp\left(-\frac{3\epsilon^2(\alpha+\beta)n}{640B^4}\right). \tag{21}$$

Step 3. Additional randomization by random signs.
Let $U_1,\ldots,U_n$ be independent and uniformly distributed over the set $\{-1,1\}$ and independent of $Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'$. Because of independence and identical distribution of $Z_1,\ldots,Z_n'$ the joint distribution of $Z_1^n,Z_1'^n$ is not affected by random interchange of corresponding components in Z_1^n and $Z_1'^n$. Therefore the first probability on the right side of inequality (21) is equal to

$$\mathbf{P}\left\{\exists f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^{n}U_i(g_f(Z_i')-g_f(Z_i))\right.$$
$$\left.\geq\frac{\epsilon}{2}(\alpha+\beta)-\frac{\epsilon^2(\alpha+\beta)}{32B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\left(\frac{1}{n}\sum_{i=1}^{n}(g_f^2(Z_i)+g_f^2(Z_i'))\right)\right\}$$

and this in turn, by the union bound, is bounded by

$$\mathbf{P}\left\{\exists f\in\mathcal{F}:\left|\frac{1}{n}\sum_{i=1}^{n}U_ig_f(Z_i')\right|\right.$$
$$\left.\geq\frac{1}{2}\left(\epsilon(\alpha+\beta)/2-\frac{\epsilon^2(\alpha+\beta)}{32B^2(1+\epsilon)}\right)+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g_f^2(Z_i')\right\}$$
$$+\mathbf{P}\left\{\exists f\in\mathcal{F}:\left|\frac{1}{n}\sum_{i=1}^{n}U_ig_f(Z_i)\right|\right.$$
$$\left.\geq\frac{1}{2}\left(\epsilon(\alpha+\beta)/2-\frac{\epsilon^2(\alpha+\beta)}{32B^2(1+\epsilon)}\right)+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g_f^2(Z_i)\right\}$$
$$=2\mathbf{P}\left\{\exists f\in\mathcal{F}:\left|\frac{1}{n}\sum_{i=1}^{n}U_ig_f(Z_i)\right|\right.$$
$$\left.\geq\epsilon(\alpha+\beta)/4-\frac{\epsilon^2(\alpha+\beta)}{64B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g_f^2(Z_i)\right\}. \tag{22}$$

Step 4. Conditioning and using covering.
Next we condition the probability on the right–hand side of (22) on Z_1^n, which is equivalent

to fixing $z_1, \ldots, z_n$ and considering

$$\mathbf{P}\left\{\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i g_f(z_i)\right| \right. \\ \left. \geq \epsilon(\alpha+\beta)/4 - \frac{\epsilon^2(\alpha+\beta)}{64B^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n} g_f^2(z_i)\right\}.$$

Let $\delta > 0$ and let $\mathcal{G}_\delta$ be a L_1 δ-cover of $\mathcal{G}_\mathcal{F} = \{g_f : f \in \mathcal{F}\}$ on $z_1, \ldots, z_n$. Fix $f \in \mathcal{F}$. Then there exists $g \in \mathcal{G}_\delta$ such that

$$\frac{1}{n}\sum_{i=1}^{n} |g(z_i) - g_f(z_i)| < \delta.$$

Without loosing generality we can assume $-4B^2 \leq g(z) \leq 4B^2$. This implies

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n} U_i g_f(z_i)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} U_i g(z_i) + \frac{1}{n}\sum_{i=1}^{n} U_i(g_f(z_i) - g(z_i))\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^{n} U_i g(z_i)\right| + \frac{1}{n}\sum_{i=1}^{n} |g_f(z_i) - g(z_i)| \\ &< \left|\frac{1}{n}\sum_{i=1}^{n} U_i g(z_i)\right| + \delta \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} g_f^2(z_i) &= \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) + \frac{1}{n}\sum_{i=1}^{n}(g_f^2(z_i) - g^2(z_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) + \frac{1}{n}\sum_{i=1}^{n}(g_f(z_i) + g(z_i))(g_f(z_i) - g(z_i)) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) - 8B^2\frac{1}{n}\sum_{i=1}^{n} |g_f(z_i) - g(z_i)| \\ &\geq \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) - 8B^2\delta. \end{aligned}$$

It follows

$$\begin{aligned}
&\mathbf{P}\left\{\exists f\in\mathcal{F}:\left|\frac{1}{n}\sum_{i=1}^{n}U_i g_f(z_i)\right|\right.\\
&\qquad\left.\geq \epsilon(\alpha+\beta)/4-\frac{\epsilon^2(\alpha+\beta)}{64B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g_f^2(z_i)\right\}\\
&\leq \mathbf{P}\left\{\exists g\in\mathcal{G}_\delta:\left|\frac{1}{n}\sum_{i=1}^{n}U_i g(z_i)\right|+\delta\right.\\
&\qquad\geq \epsilon(\alpha+\beta)/4-\frac{\epsilon^2(\alpha+\beta)}{64B^2(1+\epsilon)}\\
&\qquad\qquad\left.+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\left(\frac{1}{n}\sum_{i=1}^{n}g^2(z_i)-8B^2\delta\right)\right\}\\
&\leq |\mathcal{G}_\delta|\max_{g\in\mathcal{G}_\delta}\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i g(z_i)\right|\right.\\
&\qquad\geq \epsilon(\alpha+\beta)/4-\frac{\epsilon^2(\alpha+\beta)}{64B^2(1+\epsilon)}-\delta-\delta\frac{\epsilon(1-\epsilon)}{8(1+\epsilon)}\\
&\qquad\qquad\left.+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g^2(z_i)\right\}.
\end{aligned}$$

Next we set $\delta=\epsilon\beta/5$. This together with $B\geq 1$ and $0<\epsilon\leq\frac{1}{2}$ implies

$$\begin{aligned}
&\frac{\epsilon\beta}{4}-\frac{\epsilon^2\beta}{64B^2(1+\epsilon)}-\delta-\delta\frac{\epsilon(1-\epsilon)}{8(1+\epsilon)}\\
&=\frac{\epsilon\beta}{20}-\frac{\epsilon^2\beta}{64B^2(1+\epsilon)}-\frac{\epsilon^2(1-\epsilon)\beta}{40(1+\epsilon)}\\
&\geq 0.
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbf{P}\left\{\exists f\in\mathcal{F}:\left|\frac{1}{n}\sum_{i=1}^{n}U_i g_f(z_i)\right|\right.\\
&\qquad\left.\geq \epsilon(\alpha+\beta)/4-\frac{\epsilon^2(\alpha+\beta)}{64B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g_f^2(z_i)\right\}\\
&\leq |\mathcal{G}_{\frac{\epsilon\beta}{5}}|\max_{g\in\mathcal{G}_{\frac{\epsilon\beta}{5}}}\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i g(z_i)\right|\right.\\
&\qquad\left.\geq \frac{\epsilon\alpha}{4}-\frac{\epsilon^2\alpha}{64B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g^2(z_i)\right\}.
\end{aligned}$$

Step 5. Application of Bernstein's inequality.
In this step we use Bernstein's inequality to bound

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i g(z_i)\right|\geq\frac{\epsilon\alpha}{4}-\frac{\epsilon^2\alpha}{64B^2(1+\epsilon)}+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}g^2(z_i)\right\},$$

where $z_1, \ldots, z_n \in \mathbb{R}^d \times \mathbb{R}$ are fixed and $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ satisfies $-4B^2 \le g(z) \le 4B^2$. First we relate $\frac{1}{n}\sum_{i=1}^n g^2(z_i)$ to the variance of $U_i g(z_i)$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{V}(U_i g(z_i)) = \frac{1}{n}\sum_{i=1}^n g^2(z_i)\mathbf{V}(U_i) = \frac{1}{n}\sum_{i=1}^n g^2(z_i).$$

Thus the probability above is equal to

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n V_i\right| \ge A_1 + A_2\sigma^2\right\}$$

where

$$V_i = U_i g(z_i), \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathbf{V}(U_i g(z_i)),$$

$$A_1 = \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64B^2(1+\epsilon)}, \quad A_2 = \frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}.$$

Observe that $V_1, \ldots, V_n$ are independent random variables satisfying $|V_i| \le |g(z_i)| \le 4B^2, i = 1, \ldots, n$, and that $A_1, A_2 \ge 0$. We have by Bernstein's inequality

$$\begin{aligned}
&\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n V_i\right| \ge A_1 + A_2\sigma^2\right\} \\
&\le 2\exp\left(-\frac{n(A_1 + A_2\sigma^2)^2}{2\sigma^2 + 2(A_1 + A_2\sigma^2)\frac{8B^2}{3}}\right) \\
&= 2\exp\left(-\frac{nA_2^2}{\frac{16}{3}B^2A_2} \cdot \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8B^2A_2}\right)\sigma^2}\right) \\
&= 2\exp\left(-\frac{3 \cdot n \cdot A_2}{16B^2} \cdot \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8B^2A_2}\right)\sigma^2}\right). \qquad (23)
\end{aligned}$$

An easy calculation shows that for arbitrary $a, b, u > 0$ one has

$$\frac{(a+u)^2}{a + b \cdot u} \ge \frac{\left(a + \frac{b-2}{b}a\right)^2}{a + b\frac{b-2}{b}a} = 4a\frac{b-1}{b^2}.$$

Thus setting $a = A_1/A_2$, $b = \left(1 + \frac{3}{8B^2A_2}\right)$, $u = \sigma^2$ and using the bound above we get for the exponent in (23)

$$\begin{aligned}
\frac{3 \cdot n \cdot A_2}{16B^2} \cdot \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8B^2A_2}\right)\sigma^2} &\ge \frac{3 \cdot n \cdot A_2}{16B^2} \cdot 4 \cdot \frac{A_1}{A_2}\frac{\frac{3}{8B^2A_2}}{\left(1 + \frac{3}{8B^2A_2}\right)^2} \\
&= 18n\frac{A_1A_2}{(8B^2A_2 + 3)^2}.
\end{aligned}$$

Substituting the formulas for A_1 and A_2 and noticing

$$A_1 = \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64B^2(1+\epsilon)} \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon\alpha}{64} = \frac{15\epsilon\alpha}{64}$$

we obtain

$$\begin{aligned}
18n\frac{A_1A_2}{\left(8B^2A_2+3\right)^2} &\geq 18n\frac{15\epsilon\alpha}{64}\cdot\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\cdot\frac{1}{\left(\frac{\epsilon(1-\epsilon)}{8(1+\epsilon)}+3\right)^2} \\
&\geq 18n\frac{15\cdot\epsilon^2(1-\epsilon)\cdot\alpha}{64^2B^2(1+\epsilon)}\cdot\frac{1}{\left(\frac{1}{32}+3\right)^2} \\
&= \frac{9\cdot 15}{2\cdot 97\cdot 97}\cdot\frac{\epsilon^2(1-\epsilon)}{1+\epsilon}\cdot\frac{\alpha\cdot n}{B^2} \\
&\geq \frac{\epsilon^2(1-\epsilon)\cdot\alpha\cdot n}{140B^2(1+\epsilon)}.
\end{aligned}$$

Plugging the lower bound above into (23) we finally obtain

$$\begin{aligned}
&\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n U_i g(z_i)\right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64B^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^n g^2(z_i)\right\} \\
&\quad\leq 2\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha n}{140B^2(1+\epsilon)}\right).
\end{aligned}$$

Step 6. Bounding the covering number.
In this step we construct a L_1 $\frac{\epsilon\beta}{5}$-cover of $\{g_f : f \in \mathcal{F}\}$ on $z_1,\ldots,z_n$. Let $f_1,\ldots,f_l, l = \mathcal{N}_1(\frac{\epsilon\beta}{20B},\mathcal{F},x_1^n)$ be an $\frac{\epsilon\beta}{20B}$-cover of $\mathcal{F}$ on x_1^n. W.l.o.g. we may assume $|f_j(x)| \leq B$ for all j. Let $f \in \mathcal{F}$ be arbitrary. Then there exists an f_j such that $\frac{1}{n}\sum_{i=1}^n |f(x_i) - f_j(x_i)| < \frac{\epsilon\beta}{20B}$. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |g_f(z_i) - g_{f_j}(z_i)| \\
&= \frac{1}{n}\sum_{i=1}^n \left||f(x_i)-y_i|^2 - |m(x_i)-y_i|^2 - |f_j(x_i)-y_i|^2 + |m(x_i)-y_i|^2\right| \\
&= \frac{1}{n}\sum_{i=1}^n |f(x_i)-y_i+f_j(x_i)-y_i||f(x_i)-y_i-f_j(x_i)+y_i| \\
&\leq 4B\frac{1}{n}\sum_{i=1}^n |f(x_i)-f_j(x_i)| < \frac{\epsilon\beta}{5}.
\end{aligned}$$

Thus $g_{f_1},\ldots,g_{f_l}$ is an $\frac{\epsilon\beta}{5}$-cover of $\{g_f : f \in \mathcal{F}\}$ on z_1^n of size $\mathcal{N}_1(\frac{\epsilon\beta}{20B},\mathcal{F},x_1^n)$. Steps 3

through 6 imply

$$\begin{aligned}
&\mathbf{P}\left\{\exists f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^{n}(g_f(Z_i')-g_f(Z_i))\right.\\
&\qquad\geq\frac{\epsilon}{2}(\alpha+\beta)-\frac{\epsilon^2(\alpha+\beta)}{32B^2(1+\epsilon)}\\
&\qquad\qquad\left.+\frac{\epsilon(1-\epsilon)}{64B^2(1+\epsilon)}\frac{1}{n}\sum_{i=1}^{n}(g_f^2(Z_i)+g_f^2(Z_i'))\right\}\\
&\leq 4\sup_{x_1^n\in(\mathbb{R}^d)^n}\mathcal{N}_1\left(\frac{\epsilon\beta}{20B},\mathcal{F},x_1^n\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha n}{140B^2(1+\epsilon)}\right).
\end{aligned}$$

Step 7. Conclusion.
Steps 1, 2 and 6 imply for $n>\frac{128B^2}{\epsilon^2(\alpha+\beta)}$

$$\begin{aligned}
&\mathbf{P}\left\{\exists f\in\mathcal{F}:\mathbf{E}g_f(Z)-\frac{1}{n}\sum_{i=1}^{n}g_f(Z_i)>\epsilon(\alpha+\beta+\mathbf{E}g_f(Z))\right\}\\
&\leq\frac{32}{7}\sup_{x_1^n}\mathcal{N}_1\left(\frac{\epsilon\beta}{20B},\mathcal{F},x_1^n\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha n}{140B^2(1+\epsilon)}\right)\\
&\qquad+\frac{64}{7}\sup_{x_1^n}\mathcal{N}_1\left(\frac{\epsilon(\alpha+\beta)}{20B},\mathcal{F},x_1^n\right)\exp\left(-\frac{3\epsilon^2(\alpha+\beta)n}{640B^4}\right)\\
&\leq 14\sup_{x_1^n}\mathcal{N}_1\left(\frac{\epsilon\beta}{20B},\mathcal{F},x_1^n\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha n}{214(1+\epsilon)B^4}\right).
\end{aligned}$$

For $n\leq\frac{128B^2}{\epsilon^2(\alpha+\beta)}$ one has

$$\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha n}{214(1+\epsilon)B^4}\right)\geq\exp\left(-\frac{128}{214}\right)\geq\frac{1}{14},$$

and hence the assertion trivially follows. □

In the sequel: Derivation of upper bounds on covering numbers.

3.3 Covering numbers and VC dimension

Definition 3.2 *Let $\epsilon>0$, let $\mathcal{G}$ be a set of functions $g:\mathbb{R}^l\to\mathbb{R}$, let $1\leq p<\infty$ and let ν be a probability measure on $\mathbb{R}^l$. For $g:\mathbb{R}^l\to\mathbb{R}$ set*

$$\|g\|_{L_p(\nu)}:=\left\{\int|g(x)|^p\nu(dx)\right\}^{\frac{1}{p}}.$$

a) *A finite set of functions $g_1,\ldots,g_N\in\mathcal{G}$ with*

$$\|g_i-g_j\|_{L_p(\nu)}\geq\epsilon\quad\text{for all } 1\leq i<j\leq N$$

is called **ϵ-packing of $\mathcal{G}$** *w.r.t. $\|\cdot\|_{L_p(\nu)}$.*

b) *The* ϵ**-packing number of** $\mathcal{G}$ *w.r.t.* $\|\cdot\|_{L_p(\nu)}$

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right)$$

is defined as the cardinality of the largest ϵ*-packing of* $\mathcal{G}$ *w.r.t.* $\|\cdot\|_{L_p(\nu)}$*. Here we set* $\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right) = \infty$*, if for every* $n \in \mathbb{N}$ *there exists a* ϵ*-packing of* $\mathcal{G}$ *w.r.t.* $\|\cdot\|_{L_p(\nu)}$ *which contains* n *elements.*

c) *The* L_p**-**ϵ**-packing number of** $\mathcal{G}$ **on** z_1^n *is defined by*

$$\mathcal{M}_p\left(\epsilon, \mathcal{G}, z_1^n\right) = \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu_n)}\right),$$

where ν_n *denotes the empirical distribution corresponding to* $z_1^n = (z_1, \ldots, z_n) \in (\mathbb{R}^l)^n$*.*

Lemma 3.1 *Let* $\epsilon > 0$*, let* $\mathcal{G}$ *be a set of functions* $g : \mathbb{R}^l \to \mathbb{R}$*, let* $1 \leq p < \infty$ *and let* ν *be a probibility measure on* $\mathbb{R}^l$*. Then it holds:*

$$\mathcal{M}\left(2 \cdot \epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right) \leq \mathcal{N}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right) \leq \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right).$$

Proof. a) If $g_1, \ldots, g_N$ is a $2 \cdot \epsilon$-packing of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$, then each open ball with radius ϵ contains at most one of the g_1, $\ldots$, g_N. This shows that each ϵ-covering of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$ contains at least N functions.

b) If $g_1, \ldots, g_N$ is a ϵ-packing of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$ of maximal size, then we have that for each $g \in \mathcal{G}$

$$g_1, \ldots, g_N, g$$

is not a ϵ-packing. Consequently, for each $g \in \mathcal{G}$ there exist $j = j(g) \in \{1, \ldots, N\}$ satisfying

$$\|g - g_j\|_{L_p(\nu)} < \epsilon.$$

But this implies that $g_1, \ldots, g_N$ is a ϵ-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$. □

In order to derive upper bounds for covering numbers we consider first the special case that the functions are all indicator functions.

If $f = I_A$, $g = I_B$ for $A, B \subseteq \mathbb{R}^d$ and z_1, $\ldots$, $z_n \in \mathbb{R}^d$, then

$$\begin{aligned}\left\{\frac{1}{n}\sum_{i=1}^{n}|f(z_i) - g(z_i)|^p\right\}^{\frac{1}{p}} &\leq \max_{i=1,\ldots,n}|f(z_i) - g(z_i)| \\ &= \begin{cases} 1, & \text{if } A \cap \{z_1, \ldots, z_n\} \neq B \cap \{z_1, \ldots, z_n\} \\ 0, & \text{else.}\end{cases}\end{aligned}$$

Consequently, for $\mathcal{G} = \{1_A : A \in \mathcal{A}\}$ for $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$ and $\epsilon > 0$, we have:

$$\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n) \leq \left|\{A \cap \{z_1, \ldots, z_n\} \,:\, A \in \mathcal{A}\}\right|.$$

Definition 3.3 *Let $\mathcal{A}$ be a class of sets $A \subseteq \mathbb{R}^d$ and let $n \in \mathbb{N}$.*
a) *For $z_1, \ldots, z_n \in \mathbb{R}^d$*

$$s\left(\mathcal{A}, \{z_1, \ldots, z_n\}\right) := |\{A \cap \{z_1, \ldots, z_n\} \quad : \quad A \in \mathcal{A}\}|$$

describes the number of subsets of $\{z_1, \ldots, z_n\}$, which can be "picked out" by sets from $\mathcal{A}$.
b) *Let G be a finite subset of $\mathbb{R}^d$. We say that $\mathcal{A}$* **shatters** *G, if*

$$s(\mathcal{A}, G) = 2^{|G|},$$

i.e., if each subset of G can be represented in the form $A \cap G$ for some $A \in \mathcal{A}$.
c) *The n-***th shatter coefficient of** *$\mathcal{A}$*

$$S(\mathcal{A}, n) := \max_{z_1, \ldots, z_n \in \mathbb{R}^d} s\left(\mathcal{A}, \{z_1, \ldots, z_n\}\right)$$

is the maximal number of different subsets of n points that can be picked out by $\mathcal{A}$.

Examples: a) The set of all interval of the form $(-\infty, a]$, $a \in \mathbb{R}$, shatters all subsets of $\mathbb{R}$ of cardinality one, but it fails to shatter any subset of $\mathbb{R}$ of cardinality two (since it fails to pick out only the larger of two numbers).
b) The sets of all intervals of the form $(a, b]$, $a, b \in \mathbb{R}$, shatters all subsets of $\mathbb{R}$ of cardinality two, but it fails to shatter any subset of $\mathbb{R}$ of cardinality three.
c) The set of all half spaces in $\mathbb{R}^2$ shatter three suitably chosen points in $\mathbb{R}^2$.
d) The set of all convex sets in $\mathbb{R}^2$ shatters n (suitably chosen) points in $\mathbb{R}^2$ for any $n \in \mathbb{N}$ (choose points on a circle, and consider convex hulls of subsets of these points).

A set of sets which does not shatter a set G can not shatter any superset of G. Consequently we have:

$$S(\mathcal{A}, k) < 2^k \quad \Rightarrow \quad S(\mathcal{A}, n) < 2^n \text{ for all } n > k.$$

The largest n with $S(\mathcal{A}, n) = 2^n$ is the so–called VC dimension of $\mathcal{A}$.

Definition 3.4 *Let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$ with $\mathcal{A} \neq \emptyset$. The* **VC dimension** *(Vapnik-Chervonenkis-dimension) $V_{\mathcal{A}}$ of $\mathcal{A}$ is defined by*

$$V_{\mathcal{A}} = \sup\left\{n \in \mathbb{N} \quad : \quad S(\mathcal{A}, n) = 2^n\right\},$$

i.e., $V_{\mathcal{A}}$ is the maximal number of points, which can be shatterd by $\mathcal{A}$.

Examples: a) $\mathcal{A} = \{(-\infty, a] \ : \ a \in \mathbb{R}\} \Rightarrow V_A = 1$
b) $\mathcal{A} = \{(a, b] \ : \ a, b \in \mathbb{R}\} \Rightarrow V_A = 2$
c) $\mathcal{A} = \{A \ : \ A \text{ konvex}\} \Rightarrow V_A = \infty$

Our next theorem implies:
Either we have $S(\mathcal{A}, n) = 2^n$ for all $n \in \mathbb{N}$, or $S(\mathcal{A}, n)$ is bounded by some polynomial in n of degree $V_{\mathcal{A}}$.

Theorem 3.4 *(Vapnik and Chervonenkis (1971)).*
Let $\mathcal{A}$ be a set of subsets of $\mathbb{R}^d$ with VC dimension $V_{\mathcal{A}}$. Then we have for all $n \in \mathbb{N}$:

$$S(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Corollary 3.1 *Let $\mathcal{A}$ be a set of subsets of $\mathbb{R}^d$ with VC dimension $V_{\mathcal{A}}$.*
a)

$$S(\mathcal{A}, n) \leq (n+1)^{V_{\mathcal{A}}} \quad \textit{for all } n \in \mathbb{N}.$$

b)

$$S(\mathcal{A}, n) \leq \left(\frac{e \cdot n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}} \quad \textit{for all } n \geq V_{\mathcal{A}}.$$

Proof: a) By Theorem 3.4 and the binomial theorem it holds:

$$\begin{aligned} S(\mathcal{A}, n) &\leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i} = \sum_{i=0}^{V_{\mathcal{A}}} n \cdot (n-1) \cdot \dots \cdot (n-i+1) \cdot \frac{1}{i!} \\ &\leq \sum_{i=0}^{V_{\mathcal{A}}} n^i \cdot \frac{V_{\mathcal{A}}!}{(V_{\mathcal{A}}-i)!} \cdot \frac{1}{i!} \\ &= \sum_{i=0}^{V_{\mathcal{A}}} n^i \cdot \binom{V_{\mathcal{A}}}{i} = (n+1)^{V_{\mathcal{A}}}. \end{aligned}$$

b) If $V_{\mathcal{A}}/n \leq 1$, then Theorem 3.4 implies:

$$\begin{aligned} \left(\frac{V_{\mathcal{A}}}{n}\right)^{V_{\mathcal{A}}} \cdot S(\mathcal{A}, n) &\leq \sum_{i=0}^{V_{\mathcal{A}}} \left(\frac{V_{\mathcal{A}}}{n}\right)^{V_{\mathcal{A}}} \cdot \binom{n}{i} \\ &\leq \sum_{i=0}^{n} \left(\frac{V_{\mathcal{A}}}{n}\right)^{i} \cdot \binom{n}{i} \\ &= \left(1 + \frac{V_{\mathcal{A}}}{n}\right)^{n} \leq e^{V_{\mathcal{A}}}, \end{aligned}$$

where the last inequality follows from $1 + x \leq e^x$ $(x \in \mathbb{R})$. Consequently,

$$S(\mathcal{A}, n) \leq \left(\frac{n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}} \cdot e^{V_{\mathcal{A}}} = \left(\frac{e \cdot n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}}.$$

□

Proof of Theorem 3.4: W.l.o.g. we can assume $V_{\mathcal{A}} < n$, because otherwise the right-hand side is equal to 2^n and thus trivially greater or equal to the left-hand side.
Let $z_1, \dots, z_n \in \mathbb{R}^d$. In the sequel we show:

$$|\{A \cap \{z_1, \dots, z_n\} \quad : \quad A \in \mathcal{A}\}| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

To show this, let $F_1, \ldots, F_k$ with $k = \binom{n}{V_A+1}$ be all subsets of $\{z_1, \ldots, z_n\}$ of size (V_A+1). The definition of $V_{\mathcal{A}}$ implies that for each $i \in \{1, \ldots, k\}$ there exists $H_i \subseteq F_i$ such that

$$A \cap F_i \neq H_i \quad \text{for all } A \in \mathcal{A}$$

(since $\mathcal{A}$ does not shatter F_i because of $|F_i| > V_{\mathcal{A}}$).
$H_i \subseteq F_i \subseteq \{z_1, \ldots, z_n\}$ implies

$$(A \cap \{z_1, \ldots, z_n\}) \cap F_i \neq H_i \quad \text{for all } A \in \mathcal{A}.$$

Consequently,

$$\begin{aligned} &\{A \cap \{z_1, \ldots, z_n\} \quad : \quad A \in \mathcal{A}\} \\ &\subseteq \{C \subseteq \{z_1, \ldots, z_n\} \quad : \quad C \cap F_i \neq H_i \text{ for all } i \in \{1, \ldots, k\}\} =: \mathcal{C}_0. \end{aligned}$$

Hence it suffices to show:

$$|\mathcal{C}_0| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

This is easy in case $H_i = F_i$ for all $i \in \{1, \ldots, k\}$. Because $F_1, \ldots, F_k$ are all subsets of size $V_{\mathcal{A}} + 1$ of $\{z_1, \ldots, z_n\}$, and for $C \subseteq \{z_1, \ldots, z_n\}$ the fact

$$C \cap F_i \neq H_i = F_i \quad \text{for all } i \in \{1, \ldots, k\},$$

implies in this case that C contains at most $V_{\mathcal{A}}$ many elements, from which we can conclude:

$$|\mathcal{C}_0| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

In the sequel we will reduce the general case to the special case just treated.
To do this we set

$$H_i' = (H_i \cup \{z_1\}) \cap F_i.$$

Because of $H_i \subseteq F_i$ we augment H_i in the case $z_1 \in F_i$ and $z_1 \notin H_i$ by z_1, and otherwise H_i does not change at all (so if H_i changes, we have $z_1 \notin H_i$).
Define

$$\mathcal{C}_1 := \{C \subseteq \{z_1, \ldots, z_n\} \quad : \quad C \cap F_i \neq H_i' \text{ for all } i \in \{1, \ldots, k\}\}.$$

We show next

$$|\mathcal{C}_0| \leq |\mathcal{C}_1|. \tag{24}$$

Therefore it suffices to show

$$|\mathcal{C}_0 \setminus \mathcal{C}_1| \leq |\mathcal{C}_1 \setminus \mathcal{C}_0|,$$

which we show by proving that the mapping

$$f : \mathcal{C}_0 \setminus \mathcal{C}_1 \to \mathcal{C}_1 \setminus \mathcal{C}_0, \quad f(C) = C \setminus \{z_1\}$$

is well defined and one-to-one.
Let $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$. Then

$$C \cap F_i \neq H_i \text{ for all } i \in \{1, \ldots, k\}$$

and

$$C \cap F_{i_0} = H'_{i_0} \text{ for some } i_0 \in \{1, \ldots, k\}.$$

Consequently we have for some $i_0 \in \{1, \ldots, k\}$:

$$H'_{i_0} = C \cap F_{i_0} \neq H_{i_0}.$$

By definition of H'_i, this set differs from H_i at most by z_1, hence we can conclude

$$z_1 \in H'_{i_0} = C \cap F_{i_0} \subseteq C.$$

This implies that for $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$ we always have $z_1 \in C$, and consequently the above mapping is one-to-one, provided it is well defined.
So it remains to show that f is well defined, i.e., for all $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$ the relation

$$C \setminus \{z_1\} \in \mathcal{C}_1 \setminus \mathcal{C}_0$$

holds.
We have:

1. As seen above, $C \in \mathcal{C}_0 \setminus \mathcal{C}_1$ implies $H'_{i_0} = H_{i_0} \cup \{z_1\}$, $z_1 \notin H_{i_0}$ and $C \cap F_{i_0} = H'_{i_0}$, thus
$$C \setminus \{z_1\} \cap F_{i_0} = (C \cap F_{i_0}) \setminus \{z_1\} = H'_{i_0} \setminus \{z_1\} = H_{i_0}.$$
This shows $C \setminus \{z_1\} \notin \mathcal{C}_0$.

2. In case $z_1 \notin F_i$, we have $H_i = H'_i$, and because of $C \in \mathcal{C}_0$ we can conclude
$$(C \setminus \{z_1\}) \cap F_i = C \cap F_i \neq H_i = H'_i.$$
In case $z_1 \in F_i$, we have $z_1 \in H'_i$, which implies
$$C \setminus \{z_1\} \cap F_i \neq H'_i,$$
since the left-hand side does not contain z_1, and the right-hand side contains z_1.
Consequently we have in both cases: $C \setminus \{z_1\} \in \mathcal{C}_1$.

This proves (24).
By augmenting in the same way H'_i by $z_2, z_3, \ldots, z_n$, we get

$$|\mathcal{C}_0| \leq |\mathcal{C}_1| \leq \cdots \leq |\mathcal{C}_n|,$$

and for $\mathcal{C}_n$ all the sets $H_i^{(n)}$ satisfy the conditions of the special case above, which implies the assertion. $\square$

In order to upper bound the packing number of a set $\mathcal{G}$ of functions $g : \mathbb{R}^l \to \mathbb{R}$, it is helpful to consider the VC dimension of the set

$$\mathcal{G}^+ := \left\{ \left\{ (z,t) \in \mathbb{R}^l \times \mathbb{R} \,:\, t \le g(z) \right\} \quad : \quad g \in \mathcal{G} \right\}$$

of all subgraphs of functions of $\mathcal{G}$ helpful as can be seen from our next result.

Theorem 3.5 *Let $l \in \mathbb{N}$, $B > 0$ and let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^l \to [0,B]$ with $V_{\mathcal{G}^+} \ge 2$. Then we have for any probability measure ν on $\mathbb{R}^l$ and any $0 < \epsilon < B/4$:*

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \le 3 \cdot \left(\frac{2 \cdot e \cdot B}{\epsilon} \cdot \log \frac{3 \cdot e \cdot B}{\epsilon} \right)^{V_{\mathcal{G}^+}}.$$

Proof. We will show

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \le 3 \cdot S\left(\mathcal{G}^+, \left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) \right\rfloor \right). \tag{25}$$

This implies the assertion, because in case

$$\left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) \right\rfloor < V_{\mathcal{G}^+}$$

the assertion trivially holds, because in this case we have

$$\log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) < \frac{\epsilon}{B} \cdot (V_{\mathcal{G}^+} + 1) < \frac{\epsilon}{B} \cdot 2V_{\mathcal{G}^+} \le V_{\mathcal{G}^+},$$

and in case

$$\left\lfloor \frac{B}{\epsilon} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) \right\rfloor \ge V_{\mathcal{G}^+}$$

we can conclude by Corollary 3.1 b) that (25) implies

$$\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right) \le 3 \cdot \left(\frac{e \cdot B}{\epsilon \cdot V_{G^+}} \cdot \log\left(2 \cdot \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}\right)\right) \right)^{V_{\mathcal{G}^+}}.$$

From the last inequality we get the assertion of Theorem 3.5 by using the elementary relation

$$x \le 3 \cdot \left(\frac{a}{b} \cdot \log(2 \cdot x) \right)^b \quad \Longrightarrow \quad x \le 3 \cdot (2 \cdot a \cdot \log(3 \cdot a))^b.$$

Next we prove the above elementary relation. Let $a \in \mathbb{R}^+, b \in \mathbb{N}$ with $a \ge e$ and $b \ge 2$. We will show that

$$x \le 3 \left\{ \frac{a}{b} \log(2x) \right\}^b$$

implies

$$x < 3(2a \log(3a))^b. \tag{26}$$

Note that

$$x \leq 3\left\{\frac{a}{b}\log(2x)\right\}^b$$

is equivalent to

$$(2x)^{1/b} \leq 6^{1/b}\frac{a}{b}\log(2x) = 6^{1/b}a\log((2x)^{1/b}).$$

Set $u = (2x)^{1/b}$ and $c = 6^{1/b}a$. Then $e \leq a \leq c$ and the last inequality can be rewritten

$$u \leq c\log(u). \tag{27}$$

We will show momentarily that this implies

$$u \leq 2c\log(c). \tag{28}$$

From (28) one easily concludes (26)

$$x = \frac{1}{2}u^b \leq \frac{1}{2}(2c\log c)^b = \frac{1}{2}(2\cdot 6^{1/b}a\log(6^{1/b}a))^b \leq 3(2a\log(3a))^b,$$

where the last inequality follows from $6^{1/b} \leq 3$ for $b \geq 2$.
In conclusion we will show that (27) implies (28). Set $f_1(u) = u$ and $f_2(u) = c\log(u)$. Then it suffices to show

$$f_1(u) > f_2(u)$$

for $u > 2c\log(c)$. Because

$$f_1'(u) = 1 \geq \frac{1}{2\log(e)} \geq \frac{1}{2\log(c)} = \frac{c}{2c\log(c)} \geq \frac{c}{u} = f_2'(u)$$

for $u > 2c\log(c)$, this is equivalent to

$$f_1(2c\log(c)) > f_2(2c\log(c)).$$

This in turn is equivalent to

$$\begin{aligned} 2c\log(c) > c\log(2c\log(c)) &\Leftrightarrow 2c\log(c) > c\log(2) + c\log(c) + c\log(\log(c)) \\ &\Leftrightarrow c\log(c) - c\log(2) - c\log(\log(c)) > 0 \\ &\Leftrightarrow \log\left(\frac{c}{2\log(c)}\right) > 0 \\ &\Leftrightarrow c > 2\log(c). \end{aligned} \tag{29}$$

Set $g_1(v) = v$ and $g_2(v) = 2\log(v)$. Then

$$g_1(e) = e > 2\log(e) = g_2(e)$$

and for $v \geq e$ one has

$$g_1'(v) = 1 \geq \frac{2}{v} = g_2'(v).$$

This proves

$$g_1(v) > g_2(v)$$

for $v \geq e$, which together with $c \geq e$ implies (29).
In order to prove (25) we choose

$$\bar{\mathcal{G}} = \{g_1, \ldots, g_m\}$$

as ϵ-packing of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_1(\nu)}$ with maximal size

$$m = \mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}.\right)$$

(In fact it suffices to show (25) with $\mathcal{M}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}.\right)$ replaced by the size m of a packing for an arbitrary packing, which shows that the case of an infinite packing cannot occur). Let $Q_1, \ldots, Q_k, T_1, \ldots, T_k$ be independent random variables with $Q_1, \ldots, Q_k$ identically distributed with distribution ν and $T_1, \ldots, T_k$ identically uniformly on $[0, B]$. We set

$$\begin{aligned} R_i &= (Q_i, T_i) \quad (i = 1, \ldots, k) \\ R_1^k &= (R_1, \ldots, R_k) \end{aligned}$$

and

$$G_f = \{(z,t) : t \leq f(z)\} \quad \text{for } f \in \mathcal{G}.$$

Then we have (where the first equality follows from the definition of the shatter coefficient):

$$\begin{aligned} & S(\mathcal{G}^+, k) \\ & \geq \mathbf{E}\left\{s\left(\mathcal{G}^+, R_1^k\right)\right\} \\ & \geq \mathbf{E}\left\{s\left(\{G_f : f \in \bar{\mathcal{G}}\}, R_1^k\right)\right\} \\ & \geq \mathbf{E}\left\{s\left(\left\{G_f : f \in \bar{\mathcal{G}} \text{ and } G_f \cap R_1^k \neq G_g \cap R_1^k \text{ for all } g \in \bar{\mathcal{G}} \setminus \{f\}\right\}, R_1^k\right)\right\} \\ & = \mathbf{E}\left\{\sum_{f \in \bar{\mathcal{G}}} I_{\left\{G_f \cap R_1^k \neq G_g \cap R_1^k \text{ for all } g \in \bar{\mathcal{G}} \setminus \{f\}\right\}}\right\} \\ & = \sum_{f \in \bar{\mathcal{G}}} \mathbf{P}\left\{G_f \cap R_1^k \neq G_g \cap R_1^k \text{ for all } g \in \bar{\mathcal{G}} \setminus \{f\}\right\} \\ & = \sum_{f \in \bar{\mathcal{G}}} \left(1 - \mathbf{P}\left\{\exists g \in \bar{\mathcal{G}} \setminus \{f\} : G_f \cap R_1^k = G_g \cap R_1^k\right\}\right) \\ & \geq \sum_{f \in \bar{\mathcal{G}}} \left(1 - m \cdot \max_{g \in \bar{\mathcal{G}} \setminus \{f\}} \mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\}\right). \end{aligned}$$

For arbitrary $f, g \in \bar{\mathcal{G}}$ with $f \neq g$ we have because of $R_1, \ldots, R_k$ independent and identically distributed

$$\begin{aligned} & \mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\} \\ & = \mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}, \ldots, G_f \cap \{R_k\} = G_g \cap \{R_k\}\right\} \\ & = \left(\mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}\right\}\right)^k. \end{aligned}$$

T_1 uniformly distributed on $[0,B]$, $g(Q_1), f(Q_1) \in [0,B]$, the choice of Q_1 and $\bar{\mathcal{G}}$ ϵ-packing w.r.t. $\|\cdot\|_{L_1(\nu)}$ imply

$$\begin{aligned}
&\mathbf{P}\left\{G_f \cap \{R_1\} = G_g \cap \{R_1\}\right\} \\
&= 1 - \mathbf{P}\left\{G_f \cap \{R_1\} \neq G_g \cap \{R_1\}\right\} \\
&= 1 - \mathbf{E}\left\{\mathbf{P}\left\{G_f \cap \{R_1\} \neq G_g \cap \{R_1\} \middle| Q_1\right\}\right\} \\
&= 1 - \mathbf{E}\left\{\mathbf{P}\left\{g(Q_1) < T_1 \leq f(Q_1) \text{ or } f(Q_1) < T_1 \leq g(Q_1) \middle| Q_1\right\}\right\} \\
&= 1 - \mathbf{E}\left\{\frac{|f(Q_1) - g(Q_1)|}{B}\right\} \\
&= 1 - \frac{1}{B}\int |f(x) - g(x)| \nu(dx) \\
&\leq 1 - \frac{\epsilon}{B}.
\end{aligned}$$

Using $1 + x \leq e^x$ $(x \in \mathbb{R})$ we can conclude

$$\mathbf{P}\left\{G_f \cap R_1^k = G_g \cap R_1^k\right\} \leq \left(1 - \frac{\epsilon}{B}\right)^k \leq \exp\left(-\frac{\epsilon \cdot k}{B}\right),$$

from which we get by using the lower bound on $S(\mathcal{G}^+, k)$ derived above

$$S(\mathcal{G}^+, k) \geq m \cdot \left(1 - m \cdot \exp\left(-\frac{\epsilon \cdot k}{B}\right)\right).$$

Next we set

$$k = \lfloor \frac{B}{\epsilon} \cdot \log(2 \cdot m) \rfloor.$$

Then

$$\begin{aligned}
&1 - m \cdot \exp\left(-\frac{\epsilon \cdot k}{B}\right) \\
&\geq 1 - m \cdot \exp\left(-\frac{\epsilon}{B} \cdot \left(\frac{B}{\epsilon} \cdot \log(2 \cdot m) - 1\right)\right) \\
&= 1 - m \cdot \frac{1}{2m} \cdot \exp\left(\frac{\epsilon}{B}\right) \\
&= 1 - \frac{1}{2} \cdot \exp\left(\frac{\epsilon}{B}\right) \\
&\geq 1 - \frac{1}{2} \cdot \exp\left(\frac{1}{4}\right) \geq \frac{1}{3}
\end{aligned}$$

and hence

$$S\left(\mathcal{G}^+, \lfloor \frac{B}{\epsilon} \cdot \log(2 \cdot m) \rfloor\right) \geq \frac{1}{3} \cdot m,$$

which proves (25). □

Any application of Theorem 3.5 requires a bound on $V_{\mathcal{G}^+}$. Our next result contains such a bound.

Theorem 3.6 *Let $\mathcal{G}$ be a r-dimensional vector space of real functions on $\mathbb{R}^d$, and set*

$$\mathcal{A} = \{\{z \,:\, g(z) \geq 0\} \quad : \quad g \in \mathcal{G}\}\,.$$

Then

$$V_{\mathcal{A}} \leq r.$$

If $\mathcal{G}$ satisfies the assumptions of Theorem 3.6, then

$$\begin{aligned} \mathcal{G}^+ &= \big\{\big\{(z,t) \in \mathbb{R}^d \times \mathbb{R} \quad : \quad t \leq g(z)\big\} \quad : \quad g \in \mathcal{G}\big\} \\ &\subseteq \big\{\big\{(z,t) \in \mathbb{R}^d \times \mathbb{R} \quad : \quad g(z) + \alpha \cdot t \geq 0\big\} \quad : \quad g \in \mathcal{G},\, \alpha \in \mathbb{R}\big\} \end{aligned}$$

and by Theorem 3.6 we get

$$V_{\mathcal{G}^+} \leq r + 1.$$

Proof of Theorem 3.6: Let $z_1, \ldots, z_{r+1}$ be $(r+1)$ distinct points of $\mathbb{R}^d$. We will show that

$$\{\{z \,:\, g(z) \geq 0\} \quad : \quad g \in \mathcal{G}\}$$

does not shatter these points.
To do this, we set

$$L : \mathcal{G} \to \mathbb{R}^{r+1}, \quad L(g) = (g(z_1), \ldots, g(z_{r+1}))^T\,.$$

Then L is a linear mapping, and the image $L\mathcal{G}$ of the r-dimensional vector space $\mathcal{G}$ is a subspace of dimension less than or equal to r of $\mathbb{R}^{r+1}$. Hence there exists a nonzero vector that is orthogonal to $L\mathcal{G}$, i.e., there exist $\gamma_1, \ldots, \gamma_{r+1} \in \mathbb{R}^{r+1}$ with $\gamma_i \neq 0$ for some i and

$$\gamma_1 \cdot g(z_1) + \cdots + \gamma_{r+1} \cdot g(z_{r+1}) = 0 \tag{30}$$

for all $g \in \mathcal{G}$. W.l.o.g. we have $\gamma_i < 0$ for some $i \in \{1, \ldots, r+1\}$.
Assume that there exists $g \in \mathcal{G}$ such that

$$\{z \,:\, g(z) \geq 0\}$$

picks from $\{z_1, \ldots, z_{r+1}\}$ exactly those z_j with $\gamma_j \geq 0$. Then $g(z_j)$ always has the same sign as γ_j, hence it holds

$$\gamma_j \cdot g(z_j) \geq 0 \quad (j \in \{1, \ldots, r+1\}).$$

Because of

$$\gamma_i \cdot g(z_i) > 0$$

this implies

$$\gamma_1 \cdot g(z_1) + \cdots + \gamma_{r+1} \cdot g(z_{r+1}) > 0$$

which is a contradiction to (30). □

3.4 VC dimension of sets of deep neural networks

Sets of deep neural networks are nonlinear spaces in their weights, hence Theorem 3.6 cannot be used to bound their VC dimension. But for these sets the following bound holds:

Theorem 3.7 *(Bartlett et al. (2019)) Let $\sigma(z) = \max\{z, 0\}$ be the ReLU activation function, and let $\mathcal{F}$ be a set of neural networks of some fixed topology with depth L and $W \geq 2$ many (possibly nonzero) weights. Then*

$$V_{\mathcal{F}_+} \leq c_1 \cdot L \cdot W \cdot \log W$$

for some $c_1 > 0$ which does not depend on L, W, or the number of neurons in the network.

In the proof we will need the following auxiliary results.

Lemma 3.2 *Suppose $W \leq m$ and let $f_1, ..., f_m$ be polynomials of degree at most D in W variables. Define*

$$K = |\{(sgn(f_1(a)), \ldots, sgn(f_m(a))) : a \in \mathbb{R}^W\}|,$$

where

$$sgn(z) = \begin{cases} 1 & \text{if } z \geq 0, \\ 0 & \text{if } z < 0. \end{cases}$$

Then we have

$$K \leq 2 \cdot \left(\frac{2 \cdot e \cdot m \cdot D}{W}\right)^W.$$

Proof. See Theorem 8.3 in Anthony and Bartlett (1999). A sketch of the proof goes as follows:
By slightly pertubating the f_i it is poosible to show

$$K = |\{(sgn(f_1(a)), \ldots, sgn(f_m(a))) : a \in \mathbb{R}^W \setminus \cup_{i=1}^m \{\bar{a} \in \mathbb{R}^W : f_i(\bar{a}) = 0\}\}|.$$

Let $CC(A)$ be the number of connected components of $A \subseteq \mathbb{R}^W$ (where all points are connected by a continuous curve with image inside each connected component). Since inside any connected component of

$$\mathbb{R}^W \setminus \cup_{i=1}^m \{\bar{a} \in \mathbb{R}^W : f_i(\bar{a}) = 0\},$$

a sign change of any of the $f_1(a)$, ..., $f_m(a)$ is not possible, we get

$$K \leq CC(\mathbb{R}^W \setminus \cup_{i=1}^m \{\bar{a} \in \mathbb{R}^W : f_i(\bar{a}) = 0\}).$$

Now it can be shown that the right-hand side is bounded from above by

$$\sum_{S \subseteq \{1,\ldots,m\}, |S| \leq W} CC(\cap_{i \in S} \{\bar{a} \in \mathbb{R}^W : f_i(\bar{a}) = 0\}).$$

This in turn is equal to

$$\sum_{S\subseteq\{1,\ldots,m\},|S|\leq W} CC(\{\bar{a}\in\mathbb{R}^W : \sum_{i\in S} f_i(\bar{a})^2 = 0\}).$$

Since

$$\sum_{i\in S} f_i(\bar{a})^2$$

is a polynomial in W variables of degree $2D$, and since it can be shown that for any polynomial $f:\mathbb{R}^d\to\mathbb{R}$ of degree l we have

$$CC(\{a\in\mathbb{R}^d : f(a)=0\}) \leq l^{d-1}\cdot(l+2),$$

we get

$$\begin{aligned} K &\leq \sum_{S\subseteq\{1,\ldots,m\},|S|\leq W} (2D)^{W-1}\cdot(2D+2) \\ &\leq 2\cdot(2D)^W\cdot\sum_{i=0}^{W}\binom{m}{i} \leq 2\cdot(2D)^W\cdot\left(\frac{e\cdot m}{W}\right)^W, \end{aligned}$$

where the last inequality follows from

$$\left(\frac{W}{m}\right)^W\cdot\sum_{i=0}^{W}\binom{m}{i} \leq \sum_{i=0}^{m}\binom{m}{i}\cdot\left(\frac{W}{m}\right)^i \leq \left(1+\frac{W}{m}\right)^m \leq e^W.$$

□

Lemma 3.3 *Suppose that* $2^m \leq 2^L\cdot(m\cdot R/w)^w$ *for some* $R\geq 16$ *and* $m\geq w\geq L\geq 0$. *Then,*

$$m \leq L + w\cdot\log_2(2\cdot R\cdot\log_2(R)).$$

Proof. Let $R\geq 16$ and $m\geq w\geq L\geq 0$. We have to show

$$m > L + w\cdot\log_2(2\cdot R\cdot\log_2(R)) \quad\Longrightarrow\quad m > L + w\cdot\log_2\left(\frac{m\cdot R}{w}\right).$$

Set

$$f(x) = x - L - w\cdot\log_2\left(\frac{x\cdot R}{w}\right).$$

Then it suffices to show:

(I) $f(L + w\cdot\log_2(2\cdot R\cdot\log_2(R))) \geq 0$.

(II) $f'(x) > 0$ for all $x > L + w\cdot\log_2(2\cdot R\cdot\log_2(R))$.

Proof of (I): (I) means

$$L + w \cdot \log_2(2 \cdot R \cdot \log_2(R)) - L - w \cdot \log_2\left(\frac{(L + w \cdot \log_2(2 \cdot R \cdot \log_2(R))) \cdot R}{w}\right) \geq 0,$$

which is equivalent to

$$2 \cdot R \cdot \log_2(R)) \geq \frac{L + w \cdot \log_2(2 \cdot R \cdot \log_2(R))}{w} \cdot R,$$

which in turn is equivalent to

$$R^2 \geq 2^{L/w} \cdot 2 \cdot R \cdot \log_2(R)$$

or

$$\frac{R}{\log_2(R)} \geq 2 \cdot 2^{L/w}.$$

Because of $L/w \leq 1$ and $R \geq 16$ the last inequality is satisfied (since for $R \geq 16$ we have $R \geq 4 \cdot \log_2 R$).
Proof of (II): The derivative of

$$f(x) = x - L - w \cdot \frac{1}{\ln 2} \cdot \ln \frac{x \cdot R}{w}$$

is given by

$$f'(x) = 1 - \frac{w}{\ln 2} \cdot \frac{w}{x \cdot R} \cdot \frac{R}{w} = 1 - \frac{w}{\ln 2} \cdot \frac{1}{x}.$$

So (II) is implied by

$$x > \frac{w}{\ln 2}$$

for all $x > L + w \cdot \log_2(2 \cdot R \cdot \log_2(R))$, which holds since $R \geq 16$ implies

$$\log_2(2 \cdot R \cdot \log_2(R)) \geq \frac{1}{\ln 2}.$$

□

Proof of Theorem 3.7. Let $\mathcal{H}$ be the set of all functions h defined by

$$h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, \quad h(x, y) = g(x) - y$$

for some $g \in \mathcal{F}$. Let $(x_1, y_1), \ldots (x_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$ be such that

$$|\{(sgn(h(x_1, y_1)), \ldots, sgn(h(x_m, y_m))) \,:\, h \in \mathcal{H}\}| = 2^m, \tag{31}$$

which is equivalent to $\mathcal{F}^+$ shatters $(x_1, y_1), \ldots (x_m, y_m)$.
W.l.o.g. we assume $m \geq W$. It suffices to show

$$m \leq c_1 \cdot L \cdot W \cdot \log W. \tag{32}$$

To show this we partition $\mathcal{F}$ in subsets such that for each subset all

$$g(x_i) \quad (i = 1, \ldots, m)$$

are polynomials of some fixed degree and use Lemma 3.2 in order to derive an upper bound on the left-hand side of (31). This upper bound will depend polynomially on m which will enable us to conclude (32) by an application of Lemma 3.3.
Let

$$\theta \in \mathbb{R}^W$$

be the vector of all weights which determine a function $g \in \mathcal{F}$. Then we can write

$$\mathcal{F} = \{g(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R} \,:\, \theta \in \mathbb{R}^W\}.$$

In the sequel we construct a partition $\mathcal{P}_{L+1}$ of $\mathbb{R}^W$ such that for all $S \in \mathcal{P}_{L+1}$ we have that

$$g(x_1, \theta), \ldots, g(x_m, \theta)$$

(considered as functions of θ) are polynomials of degree at most $L+1$ for $\theta \in S$.
In order to construct this partition we construct recursively partions $\mathcal{P}_1, \ldots, \mathcal{P}_{L+1}$ of $\mathbb{R}^W$ such that for each $l \in \{1, \ldots, L+1\}$ and all $S \in \mathcal{P}_l$ all activations

$$a_i^{(l)}(x_j) \quad (i = 1, \ldots, r_i, j = 1, \ldots, m)$$

in level l (considered as a function of θ) are polynomials of degree at most l in θ for $\theta \in S$. Since all activations on level 1 (which are just linear combinations of the input variables) are linear polynomials as functions of θ this holds if we set $\mathcal{P}_1 = \{\mathbb{R}^W\}$.
Let $l \in \{2, \ldots, L+1\}$ and assume that for all $S \in \mathcal{P}_{l-1}$ all activations

$$a_i^{(l-1)} \quad (i = 1, \ldots, r_{l-1}, j = 1, \ldots, m) \tag{33}$$

in level $l-1$ (considered as a function of θ) are polynomials of degree at most $l-1$ in θ for $\theta \in S$. Application of Lemma 3.1 yields that (33) takes on in each set of $\mathcal{P}_{l-1}$ at most

$$2 \cdot \left(\frac{2 \cdot e \cdot m \cdot r_{l-1} \cdot (l-1)}{W} \right)^W \leq 2 \cdot (2 \cdot e \cdot m \cdot (l-1))^W$$

many different sign patterns. (Here we ignore neurons which have no nonzero weight and hence assume w.l.o.g. that $W \geq r_{l-1}$.)
If we partiton each set in $\mathcal{P}_{l-1}$ according to these sign patterns in

$$\Delta \leq 2 \cdot (2 \cdot e \cdot m \cdot L)^W$$

subsets, then on each set in the new partition all outputs of neurons in level $l-1$ are polynomials of degree at most $l-1$ (since they are one each set either equal to zero or equal to their activation). Consequently, on each set in this new partition all activations in level l are polynomials of degree at most l. We call this refined partion $\mathcal{P}_l$.
Using $\mathcal{P}_{L+1} = \mathcal{P}_L$ we have constructed a partition with

$$|\mathcal{P}_{L+1}| = \prod_{l=2}^{L} \frac{|\mathcal{P}_l|}{|\mathcal{P}_{l-1}|} \leq 2^L \cdot (2 \cdot e \cdot m \cdot L)^{W \cdot L}$$

such that for each set in this partition for all $(x, y) \in \{(x_1, y_1), \ldots, (x_m, y_m)\}$

$$g(x) \quad \text{and} \quad h(x, y) = g(x) - y$$

(considered as a function of θ) are polynomials of degree at most $L + 1$ in θ for $\theta \in S$. Using

$$\begin{aligned} &|\{(sgn(h(x_1, y_1)), \ldots, sgn(h(x_m, y_m))) \,:\, h \in \mathcal{H}\}| \\ &\leq \sum_{S \in \mathcal{P}_{L+1}} |\{(sgn(g(x_1, \theta) - y_1), \ldots, sgn(g(x_m, \theta) - y_m)) \,:\, \theta \in S\}| \end{aligned}$$

we can apply one more time Lemma 3.1 to conclude

$$\begin{aligned} 2^m &= |\{(sgn(h(x_1, y_1)), \ldots, sgn(h(x_m, y_m))) \,:\, h \in \mathcal{H}\}| \\ &\leq |\mathcal{P}_{L+1}| \cdot 2 \cdot \left(\frac{2 \cdot e \cdot m \cdot (L+1)}{W}\right)^W \\ &\leq 2^L \cdot (2 \cdot e \cdot m \cdot L)^{W \cdot L} \cdot 2 \cdot \left(\frac{2 \cdot e \cdot m \cdot (L+1)}{W}\right)^W \\ &\leq 2^L \cdot \left(\frac{2 \cdot e \cdot m \cdot W \cdot (L+1)^2}{W \cdot (L+1)}\right)^{W \cdot (L+1)}. \end{aligned}$$

Application of Lemma 3.3 with $w = W \cdot (L + 1)$ and $R = 2 \cdot e \cdot W \cdot (L + 1)^2 \geq 16$ (since $W \geq 2$) yields

$$\begin{aligned} m &\leq L + W \cdot (L+1) \cdot \log_2\left(2 \cdot 2 \cdot e \cdot W \cdot (L+1)^2 \cdot \log_2\left(2 \cdot e \cdot W \cdot (L+1)^2\right)\right) \\ &\leq c_1 \cdot L \cdot W \cdot \log W, \end{aligned}$$

where we have used

$$L \leq W,$$

which holds w.l.o.g. because otherwise the neural network has a layer with no connection to the previous layer. □

4 Least squares neural network regression estimates

4.1 A general result

Let (X, Y), (X_1, Y_1), (X_2, Y_2), ... be i.i.d. $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\{Y^2\} < \infty$, let

$$m : \mathbb{R}^d \to \mathbb{R}, \quad m(x) = \mathbf{E}\{Y|X = x\}$$

be the corresponding regression function, and let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. Set

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}.$$

In the sequel we use the results from Chapter 3 to derive a bound on the expected L_2 error of the least squares estimates

$$m_n(x) = m_n(x, \mathcal{D}_n) = T_{\beta_n}\tilde{m}_n(x) \tag{34}$$

where

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, \mathcal{D}_n) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 \tag{35}$$

and

$$T_{\beta_n}(z) = \max\{\min\{z, \beta_n\}, -\beta_n\}$$

for $z \in \mathbb{R}$.
Our main result is the following theorem.

Theorem 4.1 *Let $\beta \geq 1$ and assume*

$$|Y| \leq \beta \quad a.s. \tag{36}$$

Set $\beta_n = \beta$ and define the estimate m_n as above. Then

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_1 \cdot \beta^4 \cdot \frac{1 + \sup_{x_1^n \in (\mathbb{R}^d)^n} \log \mathcal{N}_1\left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_n, x_1^n\right)}{n} + 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx),
\end{aligned}$$

where $T_\beta \mathcal{F}_n = \{T_\beta f : f \in \mathcal{F}_n\}$ and $(T_\beta f)(x) = T_\beta(f(x))$.

Proof. We use the error decomposition

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \mathbf{E}\left\{|m_n(X) - Y|^2 \big| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&= \Bigg(\mathbf{E}\left\{|m_n(X) - Y|^2 \big| \mathcal{D}_n\right\} - \mathbf{E}\left\{|m(X) - Y|^2\right\} \\
&\qquad\qquad -2 \cdot \left(\frac{1}{n} \sum_{i=1}^n |m_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \Bigg) \\
&\quad +2 \cdot \left(\frac{1}{n} \sum_{i=1}^n |m_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \\
&=: T_{1,n} + T_{2,n}.
\end{aligned}$$

Because of $|T_\beta z - y| \leq |z - y|$ for $|y| \leq \beta$ the definition of m_n implies

$$\begin{aligned}
T_{2,n} &\leq 2 \cdot \left(\frac{1}{n} \sum_{i=1}^n |\tilde{m}_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \\
&= 2 \cdot \left(\inf_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right)
\end{aligned}$$

which implies

$$\begin{aligned}
\mathbf{E} T_{2,n} &\leq 2 \cdot \inf_{f \in \mathcal{F}_n} \mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(X_i) - Y_i|^2 \right) \\
&= 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

Hence it suffices to show

$$\mathbf{E} T_{1,n} \leq c_1 \cdot \frac{1 + \sup_{x_1^n \in (\mathbb{R}^d)^n} \log \mathcal{N}_1 \left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_n, x_1^n \right)}{n}.$$

To show this choose $\delta_n \geq \frac{1}{n}$. Then

$$\begin{aligned}
\mathbf{E} T_{1,n} &\leq \mathbf{E}\{(T_{1,n})_+\} = \int_0^\infty \mathbf{P}\{(T_{1,n})_+ > t\} \, dt = \int_0^\infty \mathbf{P}\{T_{1,n} > t\} \, dt \\
&\leq \delta_n + \int_{\delta_n}^\infty \mathbf{P}\{T_{1,n} > t\} \, dt.
\end{aligned}$$

We have

$$
\begin{aligned}
&\mathbf{P}\{T_{1,n} > t\} \\
&\leq \mathbf{P}\Bigg\{\exists f \in \mathcal{T}_\beta \mathcal{F}_n : \mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad\qquad -2 \cdot \left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\right) > t\Bigg\} \\
&= \mathbf{P}\Bigg\{\exists f \in T_\beta \mathcal{F}_n : \mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\} \\
&\qquad\qquad - \left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\right) \\
&\qquad\qquad > \frac{1}{2} \cdot \left(\frac{t}{2} + \frac{t}{2} + \mathbf{E}\left\{|f(X)-Y|^2\right\} - \mathbf{E}\left\{|m(X)-Y|^2\right\}\right)\Bigg\}.
\end{aligned}
$$

Application of Theorem 3.2 yields for $t \geq \delta_n \geq \frac{1}{n}$

$$
\mathbf{P}\{T_{1,n} > t\} \leq 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-\frac{t \cdot n}{16 \cdot 214 \cdot 3/2 \cdot \beta^4}\right)
$$

which implies

$$
\begin{aligned}
\mathbf{E}T_{1,n} &\leq \delta_n + \int_{\delta_n}^\infty 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-\frac{t \cdot n}{16 \cdot 214 \cdot 3/2 \cdot \beta^4}\right) dt \\
&= \delta_n + \frac{16 \cdot 214 \cdot 3/2 \cdot \beta^4}{n} \cdot 14 \cdot \sup_{x_1^n} \mathcal{N}_1\left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_n, x_1^n\right) \\
&\qquad\qquad \cdot \exp\left(-\frac{\delta_n \cdot n}{16 \cdot 214 \cdot 3/2 \cdot \beta^4}\right).
\end{aligned}
$$

With

$$
\delta_n = \frac{16 \cdot 214 \cdot 3/2 \cdot \beta^4}{n} \cdot \sup_{x_1^n} \log \mathcal{N}_1\left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_n, x_1^n\right)
$$

we get the assertion. □

Remark. It can be shown that if (36) does not hold the assertion of Theorem 4.1 still holds provided

$$
\mathbf{E}\exp(c_2 \cdot Y^2) < \infty \quad \text{and} \quad \|m\|_\infty < \infty
$$

hold and we set $\beta = \beta_n = c_3 \cdot \log n$.

4.2 Rate of convergence of least squares neural network estimates

Let σ be the ReLU activation function, let $\mathcal{F}(L,r)$ be the corresponding space of neural networks with L layers and r neurons per layer, and set

$$
\mathcal{G}_n = \left\{\sum_{k=1}^{K_n} f_k \quad : \quad f_k \in \mathcal{F}(L_n, r_n) \quad (k = 1, \ldots, K_n)\right\}
$$

for some K_n, L_n, $r_n \in \mathbb{N}$. Let $\beta_n > 0$ and set

$$\tilde{m}_n(\cdot) = \arg\min_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2$$

and

$$m_n(x) = m_n(x, \mathcal{D}_n) = T_{\beta_n} \tilde{m}_n(x),$$

i.e., m_n is the truncated least squares estimate of m corresponding to $\mathcal{G}_n$.

Theorem 4.2 *Let p, C, A, $\beta > 0$, let (X, Y), (X_1, Y_1), …, (X_n, Y_n) be i.i.d. $[-A, A]^d \times [-\beta, \beta]$ valued random variables, and assume that the corresponding regression function $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)–smooth.*
Set

$$\beta_n = \beta, \quad K_n = \lceil c_1 \cdot n^{\frac{d}{2p+d}} \rceil, \quad L_n = \lceil c_2 \cdot \log n \rceil \quad \text{and} \quad r_n = c_3.$$

Then we have for c_1, c_2 and c_3 sufficiently large

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_4 \cdot (\log n)^4 \cdot n^{-\frac{2p}{2p+d}}.$$

Proof. By Theorem 4.1 we know

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_1 \cdot \beta^4 \cdot \frac{1 + \sup_{x_1^n \in (\mathbb{R}^d)^n} \log \mathcal{N}_1 \left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{G}_n, x_1^n \right)}{n} + 2 \cdot \inf_{f \in \mathcal{G}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx).
\end{aligned}$$

By Theorem 3.7 we know

$$V_{T_\beta \mathcal{G}_n^+} \leq V_{\mathcal{G}_n^+} \leq c_5 \cdot L_n \cdot K_n \cdot L_n \cdot r_n^2 \cdot \log \left(K_n \cdot L_n \cdot r_n^2 \right) \leq c_6 \cdot (\log n)^3 \cdot n^{\frac{d}{2p+d}}.$$

(Here we have used

$$V_{T_{\beta_n} \mathcal{F}^+} \leq V_{\mathcal{F}^+},$$

where $T_{\beta_n} \mathcal{F} = \{T_{\beta_n} f : f \in \mathcal{F}\}$ for a set $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$. This holds because if $T_{\beta_n} \mathcal{F}^+$ shatters $(x_1, y_1), \ldots, (x_n, y_n)$, then $|y_i| \leq \beta_n$ $(i = 1, \ldots, n)$ holds and consequently $\mathcal{F}^+$ shatters these points, too.)
Using Lemma 3.1 and Theorem 3.5 we can conclude

$$\begin{aligned}
\frac{1 + \sup_{x_1^n \in (\mathbb{R}^d)^n} \log \mathcal{N}_1 \left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{G}_n, x_1^n \right)}{n} &\leq \frac{c_6 \cdot (\log n)^3 \cdot n^{\frac{d}{2p+d}} \cdot c_7 \cdot \log n}{n} \\
&\leq c_8 \cdot (\log n)^4 \cdot n^{\frac{-2p}{2p+d}}.
\end{aligned}$$

And by Theorem 2.1 we get

$$\inf_{f \in \mathcal{G}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \left| d(m, \mathcal{G}_n, \| \cdot \|_{\infty, [-A,A]^d}) \right|^2 \leq c_9 \cdot \frac{1}{K_n^{\frac{2p}{d}}} \leq c_{10} \cdot n^{\frac{-2p}{2p+d}}.$$

□

Remark. Modifying the estimate as in the remark after the proof of Theorem 4.1 we see that the assertion also holds if (X, Y) is a $\mathbb{R}^d \times \mathbb{R}$-valued random variable with $supp(\mathbf{P}_X)$ compact,

$$\mathbf{E}\left\{e^{c_1 \cdot Y^2}\right\} < \infty$$

and m (p, C)–smooth.

4.3 Lower bounds on the rate of convergence

In this section we show that the rate of convergence in Theorem 4.2 is optimal up to the logarithmic factor $(\log n)^4$.

Definition 4.1 *Let $\mathcal{D}$ be a class of distributions of (X, Y) and let $(a_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers.*
a) $(a_n)_{n\in\mathbb{N}}$ *is called* **lower minimax rate of convergence of $\mathcal{D}$,** *if*

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = C_1 > 0.$$

b) $(a_n)_{n\in\mathbb{N}}$ *is called* **upper minimax rate of convergence of $\mathcal{D}$,** *if there exists an estimate m_n such that*

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{a_n} = C_2 < \infty.$$

c) $(a_n)_{n\in\mathbb{N}}$ *is called* **optimal minimax rate of convergence of $\mathcal{D}$,** *if $(a_n)_{n\in\mathbb{N}}$ is a lower and an upper minimax rate of convergence of $\mathcal{D}$.*

From the previous section we know: Let $p, C > 0$ and let $\mathcal{D}$ be the class of all distributions of (X, Y) such that $X \in [0, 1]^d$ *a.s.*, $\mathbf{E}\{e^{c_1 \cdot Y^2}\} < \infty$ and $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)–smooth. Then

$$\left((\log n)^4 \cdot n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}}$$

is an upper minimax rate of convergence of $\mathcal{D}$.
In the sequel we will show that

$$\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}} \tag{37}$$

is a lower minimax rate of convergence of $\mathcal{D}$, hence the rate of convergence in Theorem 4.2 cannot be improved by more than $(\log n)^4$ (in fact it can be shown that (37) is in fact the optimal minimax rate of convergence, cf. Stone (1982)).
It suffices to show that $\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}}$ is a lower minimax rate of convergence of some suitable $\tilde{\mathcal{D}} \subseteq \mathcal{D}$.

Definition 4.2 *For $p, C > 0$ let $\mathcal{D}^{(p,C)}$ be the class of all distributions of (X, Y) satisfying:*

1. *$X \sim U([0,1]^d)$*
2. *$Y = m(X) + N$ where $N \sim N(0,1)$ and X, N independent.*
3. *m (p, C)-smooth.*
4. *$|m(x)| \leq 1$ for all $x \in [0,1]^d$.*

The main result of this subsection is:

Theorem 4.3 *Let $p, C > 0$ and define $\mathcal{D}^{(p,C)}$ as above. Then*

$$\left(n^{-\frac{2p}{2p+d}}\right)_{n\in\mathbb{N}} \tag{38}$$

is a lower minimax rate of convergence of $\mathcal{D}^{(p,C)}$.

In the proof we will need:

Lemma 4.1 *Let $u \in \mathbb{R}^l$ and let C be a random variable with values in $\{-1, 1\}$ satisfying*

$$\mathbf{P}\{C = 1\} = \frac{1}{2} = \mathbf{P}\{C = -1\}.$$

Let N be an l-dimensional standard normally distributed random variable which is independent of C, i.e., $N = (N^{(1)}, \ldots, N^{(l)})$ where $N^{(1)}$, $\ldots$, $N^{(l)}$ are independent standard normally distributed real-valued random variables which are independent of C. Set

$$Z = C \cdot u + N$$

and consider the problem of predicting the value of C from the observed value of Z. Then

$$L^* := \min_{g:\mathbb{R}^l \to \{-1,1\}} \mathbf{P}\{g(Z) \neq C\} = \Phi(-\|u\|),$$

where Φ is the cdf. of $N(0,1)$.

Proof. For arbitrary $g : \mathbb{R}^l \to \{-1, 1\}$ the independence of N and C implies

$$\begin{aligned}
&\mathbf{P}\left\{g(Z) \neq C\right\} \\
&= \mathbf{P}\left\{g(C \cdot u + N) \neq C\right\} \\
&= \mathbf{P}\left\{g(C \cdot u + N) \neq C, C = 1\right\} + \mathbf{P}\left\{g(C \cdot u + N) \neq C, C = -1\right\} \\
&= \mathbf{P}\left\{g(u + N) = -1, C = 1\right\} + \mathbf{P}\left\{g(-u + N) = 1, C = -1\right\} \\
&= \mathbf{P}\left\{g(u + N) = -1\right\} \cdot \mathbf{P}\left\{C = 1\right\} + \mathbf{P}\left\{g(-u + N) = 1\right\} \cdot \mathbf{P}\left\{C = -1\right\} \\
&= \frac{1}{2} \cdot \mathbf{P}\left\{g(u + N) = -1\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{g(-u + N) = 1\right\}.
\end{aligned}$$

Let φ be the density of N, i.e., for $v = (v^{(1)}, \ldots, v^{(l)})$ we have

$$\varphi(v) = \prod_{i=1}^{l} \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{|v^{(i)}|^2}{2}} = (2 \cdot \pi)^{-l/2} \cdot e^{-\|v\|^2/2}.$$

$u + N$ has the density $\varphi(v - u)$, and $-u + N$ has the density $\varphi(v + u)$ (which can be shown by derivating the corresponding cdf).
Hence

$$\begin{aligned}
&\mathbf{P}\{g(Z) \neq C\} \\
&= \frac{1}{2} \cdot \int I_{\{g(z)=-1\}} \cdot \varphi(z - u)\, dz + \frac{1}{2} \cdot \int I_{\{g(z)=1\}} \cdot \varphi(z + u)\, dz \\
&= \frac{1}{2} \cdot \int \left(I_{\{g(z)=-1\}} \cdot \varphi(z - u) + I_{\{g(z)=1\}} \cdot \varphi(z + u) \right) dz.
\end{aligned}$$

The right-hand side above is minimal for

$$g^*(z) = \begin{cases} 1, & \text{if } \varphi(z - u) > \varphi(z + u), \\ -1, & \text{else.} \end{cases}$$

Because of

$$\begin{aligned}
\varphi(z - u) > \varphi(z + u) &\Leftrightarrow (2 \cdot \pi)^{-l/2} \cdot e^{-\|z-u\|^2/2} > (2 \cdot \pi)^{-l/2} \cdot e^{-\|z+u\|^2/2} \\
&\Leftrightarrow \|z + u\|^2 > \|z - u\|^2 \\
&\Leftrightarrow < z, u > \quad > 0
\end{aligned}$$

we get

$$g^*(z) = \begin{cases} 1, & \text{if } < z, u > \quad > 0, \\ -1, & \text{else} \end{cases}$$

and as above we get

$$\begin{aligned}
L^* &= \mathbf{P}\{g^*(Z) \neq C\} \\
&= \mathbf{P}\{g^*(Cu + N) \neq C, C = 1\} + \mathbf{P}\{g^*(Cu + N) \neq C, C = -1\} \\
&= \frac{1}{2} \cdot \mathbf{P}\{g^*(u + N) = -1\} + \frac{1}{2} \cdot \mathbf{P}\{g^*(-u + N) = 1\} \\
&= \frac{1}{2} \cdot \mathbf{P}\{< u + N, u > \quad \leq 0\} + \frac{1}{2} \cdot \mathbf{P}\{< -u + N, u > \quad > 0\} \\
&= \frac{1}{2} \cdot \mathbf{P}\left\{\|u\|^2 + < u, N > \quad \leq 0\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{-\|u\|^2 + < u, N > \quad > 0\right\} \\
&= \frac{1}{2} \cdot \mathbf{P}\left\{< u, N > \quad \leq -\|u\|^2\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{< u, N > \quad > \|u\|^2\right\}.
\end{aligned}$$

In case $u = 0$ we have

$$L^* = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2} = \Phi(-\|u\|).$$

In case $\|u\| \neq 0$ we have that

$$< \frac{u}{\|u\|}, N >$$

is as a linear combination of independent normally distributed random variables normally distributed, and because of $\mathbf{E}\{< \frac{u}{\|u\|}, N >\} = 0$ and $\mathbf{V}\{< \frac{u}{\|u\|}, N >\} = \|u\|^2/\|u\|^2 = 1$ this random variable is standard normally distributed. Hence

$$\begin{aligned} L^* &= \frac{1}{2} \cdot \mathbf{P}\left\{< \frac{u}{\|u\|}, N > \quad \leq -\|u\|\right\} + \frac{1}{2} \cdot \mathbf{P}\left\{< \frac{u}{\|u\|}, N > \quad > \|u\|\right\} \\ &= \frac{1}{2} \cdot \Phi(-\|u\|) + \frac{1}{2} \cdot (1 - \Phi(\|u\|)) \\ &= \Phi(-\|u\|). \end{aligned}$$

□

Proof of Theorem 4.3: We prove Theorem 4.3 only in case $d = 1$, the general case will be considered in the practising course.

1. Step: Depending on n we define a subclass of $\mathcal{D}^{(p,C)}$.

Let $p = k + \beta$ with $k \in \mathbb{N}_0$ and $\beta \in (0, 1]$ and set

$$M_n = \lceil (C^2 \cdot n)^{\frac{1}{2p+1}} \rceil$$

(where $\lceil x \rceil = \inf\{z \in \mathbb{Z} : z \geq x\}$) and partition $[0, 1]$ into M_n intervals $A_{n,j}$ of length $1/M_n$. Let $a_{n,j}$ be the center of $A_{n,j}$.

Choose a bounded function $\bar{g} : \mathbb{R} \to \mathbb{R}$ satisfying

$$supp(\bar{g}) \subseteq (-1/2, 1/2), \quad \int \bar{g}^2(x)\, dx > 0 \quad \text{and} \quad \bar{g} \ (p, 2^{\beta-1})\text{-smooth}$$

(where we can achieve the last condition by rescaling a sufficiently smooth function), and set

$$g(x) = C \cdot \bar{g}(x) \quad (x \in \mathbb{R}).$$

Then

$$supp(g) \subseteq (-1/2, 1/2), \quad \int g^2(x)\, dx = C^2 \cdot \int \bar{g}^2(x)\, dx > 0$$

and

$$g \ (p, C \cdot 2^{\beta-1})\text{-smooth.}$$

For $c_n = (c_{n,1}, \ldots, c_{n,M_n}) \in \{-1, 1\}^{M_n} =: \mathcal{C}_n$ set

$$m^{(c_n)}(x) = \sum_{j=1}^{M_n} c_{n,j} \cdot g_{n,j}(x)$$

where

$$g_{n,j}(x) = M_n^{-p} \cdot g(M_n(x - a_{n,j})).$$

Then $m^{(c_n)}$ is (p, C)-smooth, which we can show as follows:

(i) For $x, z \in A_{n,i}$ we have

$$\begin{aligned}
&\left| \left(\frac{d}{dx}\right)^k m^{(c_n)}(x) - \left(\frac{d}{dx}\right)^k m^{(c_n)}(z) \right| \\
&= |c_{n,i}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) - \left(\frac{d}{dx}\right)^k g_{n,i}(z) \right| \\
&= 1 \cdot M_n^{-p} \cdot M_n^k \cdot C \cdot 2^{\beta-1} \left| M_n(x - a_{n,i}) - M_n(z - a_{n,i}) \right|^\beta \\
&\leq C \cdot 2^{\beta-1} \cdot |x - z|^\beta \leq C \cdot |x - z|^\beta.
\end{aligned}$$

(This also shows that $g_{n,i}$ is (p, C)–smooth on whole $\mathbb{R}$.)

(ii) For $x \in A_{n,i}$ and $z \in A_{n,j}$ with $i \neq j$ let $\tilde{x}$ and $\tilde{z}$, resp. be points on the border of $A_{n,i}$ bzw. $A_{n,j}$ in direction of z and x, resp. Since $g_{n,i}$ and $g_{n,j}$ are (p, C)-smooth (see above) and zero on the border we have

$$\left(\frac{d}{dx}\right)^k g_{n,i}(\tilde{x}) = 0 = \left(\frac{d}{dx}\right)^k g_{n,j}(\tilde{z}).$$

Using the result of step (i) we get

$$\begin{aligned}
&\left| \left(\frac{d}{dx}\right)^k m^{(c_n)}(x) - \left(\frac{d}{dx}\right)^k m^{(c_n)}(z) \right| \\
&= \left| c_{n,i} \cdot \left(\frac{d}{dx}\right)^k g_{n,i}(x) - c_{n,j} \cdot \left(\frac{d}{dx}\right)^k g_{n,j}(z) \right| \\
&\leq |c_{n,i}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) \right| + |c_{n,j}| \cdot \left| \left(\frac{d}{dx}\right)^k g_{n,j}(z) \right| \\
&= \left| \left(\frac{d}{dx}\right)^k g_{n,i}(x) - \left(\frac{d}{dx}\right)^k g_{n,i}(\tilde{x}) \right| + \left| \left(\frac{d}{dx}\right)^k g_{n,j}(z) - \left(\frac{d}{dx}\right)^k g_{n,j}(\tilde{z}) \right| \\
&\leq C \cdot 2^{\beta-1} \cdot |x - \tilde{x}|^\beta + C \cdot 2^{\beta-1} \cdot |z - \tilde{z}|^\beta \\
&= C \cdot 2^\beta \cdot \left(\frac{1}{2} \cdot |x - \tilde{x}|^\beta + \frac{1}{2} \cdot |z - \tilde{z}|^\beta \right) \\
&\leq C \cdot 2^\beta \cdot \left(\frac{|x - \tilde{x}|}{2} + \frac{|z - \tilde{z}|}{2} \right)^\beta \\
&\leq C \cdot (|x - \tilde{x}| + |z - \tilde{z}|)^\beta \leq C \cdot |x - z|^\beta,
\end{aligned}$$

where the third inequality follows from the inequality of Jensen and the fact that $u \mapsto u^\beta$ is on $\mathbb{R}_+ \setminus \{0\}$ concave.

This shows that the set $\bar{\mathcal{D}}_n^{(p,C)}$ of all distributions of (X, Y) satisfying

1. $X \sim U[0, 1]$,
2. $Y = m^{(c_n)}(X) + N$ for some $c_n \in \mathcal{C}_n$ and some $N \sim N(0, 1)$, where X and N are independent

is for n sufficiently large (which ensures $\|m^{(c_n)}\|_\infty \leq 1$) a subclass of $\mathcal{D}^{(p,C)}$, and it suffices to show:

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\bar{\mathcal{D}}_n^{(p,C)}} \frac{M_n^{2p}}{C^2} \cdot \mathbf{E} \int_0^1 |m_n(x) - m^{(c_n)}(x)|^2 dx > 0. \tag{39}$$

2. Step: We use a regression estimate in order to estimate the parameter $c_n \in \mathcal{C}_n$ of a distribution $(X,Y) \in \bar{\mathcal{D}}_n^{(p,C)}$.

Let m_n be an arbitrary regression estimate. By construction we have that the supports of the $g_{n,j}$ are disjoint, which implies that $\{g_{n,j} : j \in \mathbb{N}\}$ are orthogonal in L_2. Consequently, the orthogonal projection of m_n to (the linear vector space) $\{m^{(c_n)} : c_n \in \mathbb{R}^{M_n}\}$ is given by

$$\hat{m}_n(x) = \sum_{j=1}^{M_n} \hat{c}_{n,j} \cdot g_{n,j}(x)$$

where

$$\hat{c}_{n,j} = \frac{\int_{A_{n,j}} m_n(x) \cdot g_{n,j}(x)\, dx}{\int_{A_{n,j}} g_{n,j}^2(x)\, dx}.$$

For $c_n \in \mathcal{C}_n$ we have

$$\begin{aligned}
&\int_0^1 |m_n(x) - m^{(c_n)}(x)|^2 dx \\
&\geq \int_0^1 |\hat{m}_n(x) - m^{(c_n)}(x)|^2 dx \\
&= \sum_{j=1}^{M_n} \int_{A_{n,j}} |\hat{c}_{n,j} \cdot g_{n,j}(x) - c_{n,j} \cdot g_{n,j}(x)|^2 dx \\
&= \sum_{j=1}^{M_n} |\hat{c}_{n,j} - c_{n,j}|^2 \cdot \int_{A_{n,j}} g_{n,j}^2(x)\, dx \\
&= \int g^2(x)\, dx \cdot \frac{1}{M_n^{2p+1}} \cdot \sum_{j=1}^{M_n} |\hat{c}_{n,j} - c_{n,j}|^2.
\end{aligned}$$

Set

$$\tilde{c}_{n,j} = \begin{cases} 1, & \text{if } \hat{c}_{n,j} \geq 0, \\ -1, & \text{else.} \end{cases}$$

Then

$$|\hat{c}_{n,j} - c_{n,j}| \geq \frac{1}{2} \cdot |\tilde{c}_{n,j} - c_{n,j}| = I_{\{\tilde{c}_{n,j} \neq c_{n,j}\}},$$

which can be easily seen by considering the cases $\tilde{c}_{n,j} = 1$, $c_{n,j} = -1$ and $\tilde{c}_{n,j} = -1$, $c_{n,j} = 1$.

Hence

$$\int_0^1 |m_n(x) - m^{(c_n)}(x)|^2 dx \geq \int g^2(x)\, dx \cdot \frac{1}{M_n^{2p+1}} \cdot \sum_{j=1}^{M_n} I_{\{\tilde{c}_{n,j} \neq c_{n,j}\}},$$

and consequently (39) is implied by

$$\liminf_{n\to\infty} \inf_{\tilde{c}_n} \sup_{c\in\mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} > 0. \tag{40}$$

3. Step: We choose $c_n \in \mathcal{C}_n$ randomly.
Let $C_{n,1}, \ldots, C_{n,M_n}$ be independent and identically distributed random variables satisfying

$$\mathbf{P}\{C_{n,1} = 1\} = \frac{1}{2} = \mathbf{P}\{C_{n,1} = -1\},$$

which are independent of (X_1, N_1), $\ldots$, (X_n, N_n). Set

$$C_n = (C_{n,1}, \ldots, C_{n,M_n}).$$

Then

$$\begin{aligned} &\inf_{\tilde{c}_n} \sup_{c_n\in\mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} \\ &\geq \inf_{\tilde{c}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq C_{n,j}\right\}. \end{aligned}$$

Here the optimal predictor is given by

$$\bar{C}_{n,j} = \begin{cases} 1, & \text{if } \mathbf{P}\{C_{n,j} = 1 | (X_1, Y_1), \ldots, (X_n, Y_n)\} \geq \frac{1}{2}, \\ -1, & \text{else.} \end{cases}$$

Due to symmetry of the problem we have

$$\begin{aligned} \mathbf{P}\left\{\tilde{c}_{n,j} \neq C_{n,j}\right\} &= \mathbf{E}\left\{\mathbf{P}\left\{\tilde{c}_{n,j} \neq C_{n,j} | (X_1, Y_1), \ldots, (X_n, Y_n)\right\}\right\} \\ &\geq \mathbf{E}\left\{\mathbf{P}\left\{\bar{C}_{n,j} \neq C_{n,j} | (X_1, Y_1), \ldots, (X_n, Y_n)\right\}\right\} \\ &= \mathbf{P}\left\{\bar{C}_{n,j} \neq C_{n,j}\right\} = \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} \end{aligned}$$

and we get

$$\inf_{\tilde{c}_n} \sup_{c_n\in\mathcal{C}_n} \frac{1}{M_n} \sum_{j=1}^{M_n} \mathbf{P}\left\{\tilde{c}_{n,j} \neq c_{n,j}\right\} \geq \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\}.$$

Hence it suffices to show:

$$\liminf_{n\to\infty} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} > 0. \tag{41}$$

(If one does not want to use the symmetry argument above, on can also show this for each j instead just for $j = 1$).
4. Step: Proof of (41).
We use

$$\mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} = \mathbf{E}\left\{\mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1} \middle| X_1, \ldots, X_n\right\}\right\}.$$

Let $X_{i_1}, \ldots, X_{i_l}$ be those X_i satisfying $X_i \in A_{n,1}$. Then

$$(Y_{i_1}, \ldots, Y_{i_l}) = C_{n,1} \cdot (g_{n,1}(X_{i_1}), \ldots, g_{n,1}(X_{i_l})) + (N_{i_1}, \ldots, N_{i_l}) \,. \tag{42}$$

All Y_j with $X_j \notin A_{n,1}$ do depend only on $C_{n,2}, \ldots, C_{n,M_n}$ and

$$\{(X_r, N_r) : r \notin \{i_1, \ldots, i_l\}\}$$

and are consequently independent of the data in (42) given $X_1, \ldots, X_n$. If we also condition on all those random variables then we can conclude because of

$$g_{n,1}(X_j) = 0 \quad \text{for } X_j \notin A_{n,1}$$

from Lemma 4.1

$$\begin{aligned} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1} \middle| X_1, \ldots, X_n\right\} &= \Phi\left(-\sqrt{\sum_{r=1}^{l} g_{n,1}^2(X_{i_r})}\right) \\ &= \Phi\left(-\sqrt{\sum_{i=1}^{n} g_{n,1}^2(X_i)}\right), \end{aligned}$$

where Φ is the cdf. of $N(0,1)$.
It is easy to see (e.g., by computation of the second derivative) that

$$x \mapsto \Phi(-\sqrt{x})$$

is convex. Application of the inequality of Jensen yields

$$\begin{aligned} \mathbf{P}\left\{\bar{C}_{n,1} \neq C_{n,1}\right\} &= \mathbf{E}\left\{\Phi\left(-\sqrt{\sum_{i=1}^{n} g_{n,1}^2(X_i)}\right)\right\} \\ &\geq \Phi\left(-\sqrt{\mathbf{E}\left\{\sum_{i=1}^{n} g_{n,1}^2(X_i)\right\}}\right) \\ &= \Phi\left(-n \cdot \int_0^1 g_{n,1}^2(x)dx\right) \\ &= \Phi\left(-n \cdot M_n^{-(2p+1)} \cdot C^2 \int_0^1 \bar{g}^2(x)dx\right) \\ &\geq \Phi\left(-\int_0^1 \bar{g}^2(x)dx\right), \end{aligned}$$

since

$$M_n = \lceil (C^2 \cdot n)^{\frac{1}{2p+1}} \rceil \geq (C^2 \cdot n)^{\frac{1}{2p+1}}.$$

□

4.4 Deep learning as a remedy against the curse of dimensionality

The optimal rate of convergence

$$n^{-\frac{2p}{2p+d}}$$

for estimation of a (p, C)–smooth regression function gets worse in case that d (the dimension of X) is large compared to p (so-called *curse of dimensionality*). Since this rate is optimal, there is no chance to get a better rate regardedless what kind of estimate we use. The only way to circumvent this problem is to impose additional conditions on the regression function, which enable some estimates to achieve a better rate. In this section we show that deep neural network can achieve better rates in case that the high-dimensional regression function is a composition of suitable functions. This effect occurs due to the network structure of the neural network.
We consider in the sequel functions, which fulfill the following definition:

Definition 4.3 *(Kohler and Langer (2021)). Let $d \in \mathbb{N}$ and $m : \mathbb{R}^d \to \mathbb{R}$ and let $\mathcal{P}$ be a subset of $(0, \infty) \times \mathbb{N}$.*
a) *We say that m satisfies a hierarchical composition model of level 1 with order and smoothness constraint $\mathcal{P}$, if there exists $(p, K) \in \mathcal{P}$, a (p, C)–smooth function $g : \mathbb{R}^K \to \mathbb{R}$ and some $\pi : \{1, \dots, K\} \to \{1, \dots, d\}$ such that*

$$m(\mathbf{x}) = g(x^{(\pi(1))}, \dots, x^{(\pi(K))}) \quad \textit{for all } \mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top \in \mathbb{R}^d.$$

b) *We say that m satisfies a hierarchical composition model of level $l + 1$ with order and smoothness constraint $\mathcal{P}$, if there exist $(p, K) \in \mathcal{P}$, $C > 0$, $g : \mathbb{R}^K \to \mathbb{R}$ and $f_1, \dots, f_K : \mathbb{R}^d \to \mathbb{R}$, such that g is (p, C)–smooth, $f_1, \dots, f_K$ satisfy a hierarchical composition model of level l with order and smoothness constraint $\mathcal{P}$ and*

$$m(\mathbf{x}) = g(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})) \quad \textit{for all } \mathbf{x} \in \mathbb{R}^d.$$

For $l = 1$ and some order and smoothness constraint $\mathcal{P} \subseteq (0, \infty) \times \mathbb{N}$ our space of hierarchical composition models becomes

$$\begin{aligned}\mathcal{H}(1, \mathcal{P}) = \{h : \mathbb{R}^d \to \mathbb{R} : h(\mathbf{x}) = g(x^{(\pi(1))}, \dots, x^{(\pi(K))}), \text{where} \\ g : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C) \text{ –smooth for some } (p, K) \in \mathcal{P}, \\ C > 0 \text{ and } \pi : \{1, \dots, K\} \to \{1, \dots, d\}\}.\end{aligned}$$

For $l > 1$, we recursively define

$$\begin{aligned}\mathcal{H}(l, \mathcal{P}) := \{h : \mathbb{R}^d \to \mathbb{R} : h(\mathbf{x}) = g(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})), \text{where} \\ g : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C) \text{–smooth for some } (p, K) \in \mathcal{P}, \\ C > 0 \text{ and } f_i \in \mathcal{H}(l - 1, \mathcal{P})\}.\end{aligned}$$

Next we introduce sets of *sparse* neural networks, where we control the number τ of weights which are nonzero. To do this, let σ be the ReLU activation function. For $L, r, \tau \in \mathbb{N}$ let

$$\mathcal{F}_{sparse}(L, r, \tau)$$

be the set of all feedforward neural networks with activation function σ, L layers and r neurons in each layer, where at most τ of its weights are nonzero.
We consider the following truncated least suqares estimate:

$$m_n(x) = m_n(x, \mathcal{D}_n) = T_{\beta_n} \tilde{m}_n(x) \tag{43}$$

where

$$\tilde{m}_n(\cdot) = \tilde{m}_n(\cdot, \mathcal{D}_n) = \arg \min_{f \in \mathcal{F}_{sparse}(L_n, r_n, \tau_n)} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{44}$$

The main result is the following bound on its expected L_2 error.

Theorem 4.4 *(Schmidt-Hieber (2020), cf. also Bauer and Kohler (2019)).*
Let A, $\beta > 0$, let (X, Y), (X_1, Y_1), ..., (X_n, Y_n) be i.i.d. $[-A, A]^d \times [-\beta, \beta]$ valued random variables, and let $m(x) = \mathbf{E}\{Y|X = x\}$ be the corresponding regression function. Let $\mathcal{P}$ be a finite subset of $[1, \infty) \times \mathbb{N}$, let $l \in \mathbb{N}$ and assume that m satisfies a hierarchical composition model of level l with order and smoothness constraint $\mathcal{P}$.
Set

$$\beta_n = \beta, \quad L_n = \lceil c_1 \cdot \log n \rceil, \quad r_n = \lceil c_2 \cdot \max_{(p,K) \in \mathcal{P}} n^{\frac{K}{2p+K}} \rceil$$

and

$$\tau_n = \lceil c_3 \cdot (\log n) \cdot \max_{(p,K) \in \mathcal{P}} n^{\frac{K}{2p+K}} \rceil,$$

and define the least squares neural network regression estimate m_n as above.
Then we have for c_1, c_2 and c_3 sufficiently large that for any sufficiently large n

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_4 \cdot (\log n)^4 \cdot \max_{(p,K) \in \mathcal{P}} n^{-\frac{2p}{2p+K}}$$

holds.

Remark. The rate of convergence in the above theorem does not depend on the dimension d of the predictor variable X, hence the above least squares estimate is able to circumvent the curse of dimensionality in case that the regression function satisfies a suitable hierarchical composition model.

Proof. By Theorem 4.1 we know

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_5 \cdot \beta^4 \cdot \frac{1 + \sup_{x_1^n \in (\mathbb{R}^d)^n} \log \mathcal{N}_1 \left(\frac{1}{80 \cdot \beta \cdot n}, T_\beta \mathcal{F}_{sparse}(L_n, r_n, \tau_n), x_1^n \right)}{n} \\
&\qquad + 2 \cdot \inf_{f \in \mathcal{F}_{sparse}(L_n, r_n, \tau_n)} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx),
\end{aligned}$$

so it suffices to show

$$\frac{1+\sup_{x_1^n\in(\mathbb{R}^d)^n}\log\mathcal{N}_1\left(\frac{1}{80\cdot\beta\cdot n},T_\beta\mathcal{F}_{sparse}(L_n,r_n,\tau_n),x_1^n\right)}{n}\leq c_6\cdot(\log n)^4\cdot\max_{(p,K)\in\mathcal{P}}n^{-\frac{2p}{2p+K}} \tag{45}$$

and

$$\inf_{f\in\mathcal{F}_{sparse}(L_n,r_n,\tau_n)}\int|f(x)-m(x)|^2\mathbf{P}_X(dx)\leq c_6\cdot(\log n)^4\cdot\max_{(p,K)\in\mathcal{P}}n^{-\frac{2p}{2p+K}}. \tag{46}$$

If we fix the τ_n positions where the weights in $\mathcal{F}_{sparse}(L_n,r_n,\tau_n)$ are allowed to be nonzero, then the VC dimension of this function space is bounded by Theorem 3.7 by

$$c_7\cdot L_n\cdot\tau_n\cdot\log(\tau_n),$$

from which we conclude by Theorem 3.5 that the $L_1{-}\frac{1}{80\cdot\beta\cdot n}$ covering number on z_1^n of a truncated version of this function space is bounded by

$$c_9\cdot(c_{10}\cdot n)^{2\cdot c_7\cdot L_n\cdot\tau_n\cdot\log(L_n\cdot\tau_n)}.$$

Since (for n large) there are at most

$$\binom{r_n+1+L_n\cdot r_n\cdot(r_n+1)+r_n\cdot(d+1)}{\tau_n}\leq c_{11}\cdot n^{2\cdot\tau_n}$$

many possibilities to choose the positions of these weights we see that we have for n large

$$\begin{aligned}&\sup_{x_1^n\in(\mathbb{R}^d)^n}\log\mathcal{N}_1\left(\frac{1}{80\cdot\beta\cdot n},T_\beta\mathcal{F}_{sparse}(L_n,r_n,\tau_n),x_1^n\right)\\&\leq\log\left(c_{11}\cdot n^{2\cdot\tau_n}\cdot c_9\cdot(c_{10}\cdot n)^{2\cdot c_7\cdot L_n\cdot\tau_n\cdot\log(L_n\cdot\tau_n)}\right)\leq c_{12}\cdot(\log n)^3\cdot\tau_n,\end{aligned}$$

which implies

$$\begin{aligned}&\frac{1+\sup_{x_1^n\in(\mathbb{R}^d)^n}\log\mathcal{N}_1\left(\frac{1}{80\cdot\beta\cdot n},T_\beta\mathcal{F}_{sparse}(L_n,r_n,\tau_n),x_1^n\right)}{n}\\&\leq c_{13}\cdot(\log n)^3\cdot\frac{\tau_n}{n}\leq c_6\cdot(\log n)^4\cdot\max_{(p,K)\in\mathcal{P}}n^{-\frac{2p}{2p+K}}.\end{aligned}$$

In the remainder of the proof we show (46).
We observe in a first step, that one has to compute different hierarchical composition models of some level i ($i\in\{1,\dots,l-1\}$) to compute a function $h_1^{(l)}\in\mathcal{H}(l,\mathcal{P})$. Let $\tilde{N}_i$ denote the number of hierarchical composition models of level i, needed to compute $h_1^{(l)}$. We denote in the following by

$$h_j^{(i)}:\mathbb{R}^d\to\mathbb{R} \tag{47}$$

the j–th hierarchical composition model of some level i ($j\in\{1,\dots,\tilde{N}_i\}, i\in\{1,\dots,l\}$), that applies a $(p_j^{(i)},C)$–smooth function $g_j^{(i)}:\mathbb{R}^{K_j^{(i)}}\to\mathbb{R}$ with $p_j^{(i)}=q_j^{(i)}+s_j^{(i)}$, $q_j^{(i)}\in\mathbb{N}_0$

and $s_j^{(i)} \in (0,1]$, where $(p_j^{(i)}, K_j^{(i)}) \in \mathcal{P}$. The computation of $h_1^{(l)}(\mathbf{x})$ can then be recursively described as follows:

$$h_j^{(i)}(\mathbf{x}) = g_j^{(i)}\left(h^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}(\mathbf{x}), \ldots, h^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}(\mathbf{x})\right) \tag{48}$$

for $j \in \{1, \ldots, \tilde{N}_i\}$ and $i \in \{2, \ldots, l\}$ and

$$h_j^{(1)}(\mathbf{x}) = g_j^{(1)}\left(x^{\left(\pi(\sum_{t=1}^{j-1} K_t^{(1)}+1)\right)}, \ldots, x^{\left(\pi(\sum_{t=1}^{j} K_t^{(1)})\right)}\right) \tag{49}$$

for some function $\pi : \{1, \ldots, \tilde{N}_1\} \to \{1, \ldots, d\}$. Furthermore for $i \in \{1, \ldots, l-1\}$ the recursion

$$\tilde{N}_l = 1 \text{ and } \tilde{N}_i = \sum_{j=1}^{\tilde{N}_{i+1}} K_j^{(i+1)} \tag{50}$$

holds. Set

$$g_{\max} := \max\left\{ \max_{\substack{i\in\{1,\ldots,l\}\\ j\in\{1,\ldots,\tilde{N}_i\}}} \|g_j^{(i)}\|_\infty, A \right\}.$$

For the approximation of $g_j^{(i)}$ we will use the networks

$$\hat{f}_{g_j^{(i)}} \in \mathcal{G}$$

described in Theorem 2.1, where

$$K = \left\lceil c_{14} \cdot n^{\frac{1}{2\cdot p_j^{(i)}+K_j^{(i)}}} \right\rceil, L = \lceil c_{15} \cdot \log n \rceil, r = 36 \cdot \lceil p_j^{(i)} \rceil + 54 \cdot K_j^{(i)}$$

for $j \in \{1, \ldots, \tilde{N}_i\}$ and $i \in \{1, \ldots, l\}$, which satisfies

$$\|\hat{f}_{g_j^{(i)}} - g_j^{(i)}\|_{\infty, [-g_{\max}, g_{\max}]^{K_j^{(i)}}} \leq c_{17} \cdot n^{-\frac{p_j^{(i)}}{2\cdot p_j^{(i)}+K_j^{(i)}}}.$$

To compute the values of $h_1^{(1)}, \ldots, h_{\tilde{N}_1}^{(1)}$ we use the networks

$$\begin{aligned}
\hat{h}_1^{(1)}(\mathbf{x}) &= \hat{f}_{g_1^{(1)}}\left(x^{(\pi(1))}, \ldots, x^{(\pi(K_1^{(1)}))}\right)\\
&\vdots\\
\hat{h}_{\tilde{N}_1}^{(1)}(\mathbf{x}) &= \hat{f}_{g_{\tilde{N}_1}^{(1)}}\left(x^{(\pi(\sum_{t=1}^{\tilde{N}_1-1} K_t^{(1)}+1))}, \ldots, x^{(\pi(\sum_{t=1}^{\tilde{N}_1} K_t^{(1)}))}\right).
\end{aligned}$$

To compute the values of $h_1^{(i)}, \dots, h_{\tilde{N}_i}^{(i)}$ ($i \in \{2, \dots, l\}$) we use the networks

$$\hat{h}_j^{(i)}(\mathbf{x}) = \hat{f}_{g_j^{(i)}}\left(\hat{h}^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}(\mathbf{x}), \dots, \hat{h}^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}(\mathbf{x})\right)$$

for $j \in \{1, \dots, \tilde{N}_i\}$. Finally we set

$$t_1(\mathbf{x}) = \hat{h}_1^{(l)}(\mathbf{x}).$$

Since each $\hat{h}_j^{(i)}$ ($j \in \{1, \dots, \tilde{N}_i\}$) needs $\lceil c_{15} \cdot \log n \rceil$ many layers and at most

$$c_{18} \cdot \max_{(p,K)\in\mathcal{P}} n^{\frac{K}{2p+K}}$$

neurons per layer, this composed network is contained in the class

$$\mathcal{F}_{sparse}(L_n, r_n, \tau_n).$$

Next we use an induction on i to show that t_1 satisfies

$$\|t_1 - m\|_{\infty,[-A,A]^d} \leq c_{19} \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}}, \tag{51}$$

which implies (46).
Since each $g_j^{(i)}$ satisfies the assumptions of Theorem 2.1, we can conclude that

$$\left|\hat{f}_{g_j^{(i)}}(\mathbf{x}) - g_j^{(i)}(\mathbf{x})\right| \leq c_{17} \cdot n^{-\frac{p_j^{(i)}}{2\cdot p_j^{(i)}+K_j^{(i)}}} \leq c_{20} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}} \tag{52}$$

for $\mathbf{x} \in [-2g_{\max}, 2g_{\max}]^{K_j^{(i)}}$.
We show by induction that we have for all $x \in [-A, A]^d$

$$\left|\hat{h}_j^{(i)}(\mathbf{x}) - h_j^{(i)}(\mathbf{x})\right| \leq c_{17} \cdot i \cdot (K_{max} \cdot C_{Lip})^{i-1} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}}. \tag{53}$$

By (52) we can conclude that

$$\left|\hat{h}_j^{(1)}(\mathbf{x}) - h_j^{(1)}(\mathbf{x})\right| \leq c_{17} \cdot 1 \cdot (K_{max} \cdot C_{Lip})^{1-1} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}}$$

for $j \in \{1, \dots, \tilde{N}_1\}$. Assume now that (53) holds for some $i-1$ and every $j \in \{1, \dots, \tilde{N}_{i-1}\}$. Then for n sufficiently large

$$\left|\hat{h}_j^{(i-1)}(\mathbf{x})\right| \leq \left|\hat{h}_j^{(i-1)}(\mathbf{x}) - h_j^{(i-1)}(\mathbf{x})\right| + g_{\max} \leq 2 \cdot g_{\max}$$

follows directly by the induction hypothesis. Using (52) and the Lipschitz continuity of $g_j^{(i)}$ we can conclude that

$$\begin{aligned}
&\left|\hat{h}_j^{(i)}(\mathbf{x}) - h_j^{(i)}(\mathbf{x})\right| \\
&\leq \left|\hat{f}_{wide,g_j^{(i)}}\left(\hat{h}^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}, \dots, \hat{h}^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}\right) - g_j^{(i)}\left(\hat{h}^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}, \dots, \hat{h}^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}\right)\right| \\
&\quad + \left|g_j^{(i)}\left(\hat{h}^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}, \dots, \hat{h}^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}\right) - g_j^{(i)}\left(h^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}(x), \dots, h^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}(x)\right)\right| \\
&\leq c_{17} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}} + K_j^{(i)} \cdot C_{Lip} \cdot c_{17} \cdot (i-1) \cdot (K_{max} \cdot C_{Lip})^{i-2} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}} \\
&\leq c_{17} \cdot i \cdot (K_{max} \cdot C_{Lip})^{i-1} \cdot \max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}}.
\end{aligned}$$

The proof is complete. □

References

[1] Anthony, M., and Bartlett, P. L. (2009). *Neural Network Learning: Theoretical Foundations.* Cambridge University Press.

[2] Bartlett, P., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight VC-dimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research* **20**, pp. 1-17.

[3] Bauer, B., and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics* **4**, pp. 2261–2285.

[4] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York, USA.

[5] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution–Free Theory of Nonparametric Regression*. Springer.

[6] Kohler, M., and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates using ReLU activation functions. *Annals of Statistics* **49**, pp. 2231-2249. Preprint, *arXiv: 1908.11133.*

[7] Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory* **42**, pp. 2118-2132.

[8] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

[9] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Annals of Statistics* **48**, pp. 1875–1897.

[10] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040-1053.

[11] Vapnik, V. N., and Chervonenkis, A. Ya. (1971). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications* **26**, pp. 821–832.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin,I. (2017). Attention is all you need. *arXiv: 1706.03762.*

[13] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks* **94**, pp. 103–114.