Robustness Properties of Deep Neural Networks

Veenstra, Lasse

L. Veenstra

Robustness properties of deep neural networks

Bachelor thesis

July 8, 2024

Thesis supervisor: dr. A.M. Dürre

Leiden University
Mathematical Institute

Contents

1 Introduction.

Let $S = (X_i, Y_i)_{i=1}^n$ denote a sample of n independent and identically distributed (i.i.d.) observations with distribution $Z = (X, Y)$, where $Y \in \mathbb{R}$ and $X \in [0,1]^d$ for a fixed dimension $d \in \mathbb{N}_{\geq 1}$. We consider the nonparametric regression model given by:

$$Y_i = f_0(X_i) + \varepsilon_i, \tag{1.1}$$

where $f_0 : [0,1]^d \longrightarrow \mathbb{R}$ represents the unknown regression function to be estimated using the sample S and ε_i are error terms i.i.d. independent of X_i. Various parametric estimation methods, such as linear regression and logistic regression [Hosmer Jr et al., 2013] are well established, along with nonparametric methods like splines [Marsh and Cormier, 2001]. This thesis focuses on nonparametric regression using deep neural networks.

In a classical setting, it is typically assumed that $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. However, this can be insufficient in some scenarios. One such scenario involves the presence of outliers, which will be the focus of this thesis. Outliers are observations that differ significantly from the other observations, this can be due to measurement errors, but they can also be inherent to the data structure. While one approach to dealing with outliers is to detect and remove them, this strategy poses problems as it may lead to underestimation of extreme scenarios. Instead of discarding outliers, we propose to model and work with them by assuming a different distribution for ε_i. To do so, we explicitly do not require ε_i to have zero mean. In Section 3, ε_i is assumed to be sub-exponential [see Definition 3.1], while in Section 4 and Section 5, ε_i is only assumed to have a certain number of finite moments.

As previously mentioned, in this thesis, estimating the regression function f_0 will be done using deep neural networks. Many consider the Mark I, in Rosenblatt [1961], to be first implementation of a neural network. Since his work in 1961, there have been significant advances in computer hardware and software. This abundance of compute power, combined with more data, allowed the previously not so popular deep neural network to become a central tool in modern artificial intelligence. With this rise in popularity among computer scientists came a growing theoretical interest among mathematicians, especially in recent years. Some notable works are Schmidt-Hieber [2020], Shen et al. [2022], Jiao et al. [2023] and Shen et al. [2021]. All of these papers prove upper bounds on the approximation error made by neural networks. This thesis focuses in particular on the work of Jiao et al. [2023] and Shen et al. [2021]. Jiao et al. [2023] give an upper bound on the prediction error of the so called empirical risk minimizer [see Section 2.3] under the assumption that ε_i is sub-exponential and square loss. We generalize this result to a milder condition on ε_i. Furthermore, Shen et al. [2021] show similar upper bounds on the prediction error, but for a general loss function that is Lipschitz continuous.

Source	Loss function	ε assumption	Convergence rate
Jiao et al. [2023] [Lemma 3.7]	Square loss	ε is sub-exponential	$O\left(\frac{(\log n)^5}{n}\right)$
Ours [Lemma 4.1]	Square loss	$\mathbb{E}[\lvert\varepsilon\rvert^{2+\delta}] < \infty$, $\delta > 0$	$O\left(\left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}}\right)$
Shen et al. [2021] [Lemma 5.1]	Lipschitz cont. loss	$\mathbb{E}[\lvert\varepsilon\rvert^{1+\delta}] < \infty$, $\delta > 0$	$O\left(\frac{\log n}{n^{\frac{\delta}{1+\delta}}}\right)$

Table 1: An overview of convergence rates on the prediction error of the empirical risk minimizer covered in this thesis for various loss functions and assumptions on ε. In all cases it is assumed that f^* [see (2.3)] can be written as a neural network.

Some important results covered in this thesis can be best summarized in Table 1, where for each loss function and assumption on ε_i, the convergence rate of the empirical risk minimizer is given.

Furthermore, using these results we obtain that the empirical risk minimizer is a consistent estimator for f^*. From this it will follow that deep neural networks are robust against mean-zero outliers when square loss is used. At the same time, it will show that if there are outliers in the data that do not have expectation 0, a bias will created when square loss is used.

In a small simulation study in Section 6, some simple regression functions are approximated using neural networks. Then, the convergence rate is estimated and compared to what we theoretically predict.

2 Preliminaries.

In this section, we introduce ReLU neural networks along with the general problem of regression and an estimation approach. We start with the standard square loss function and show some general properties of the square loss. Then we introduce a more general loss function that is Lipschitz continuous. Finally, we show how the regression setting can incorporate the presence of outliers in the data.

2.1 Neural networks.

Definition 2.1 (Multi-layer Perceptron)**.** Consider some vector $(p_0, p_1, \dots, p_{\mathcal{D}}, p_{\mathcal{D}+1}) \in \mathbb{N}^{\mathcal{D}+2}$ for some $\mathcal{D} \in \mathbb{N}$. A *multi-layer perceptron* (MLP) is a function $f : \mathbb{R}^{p_0} \longrightarrow \mathbb{R}^{p_{\mathcal{D}+1}}$ that can be expressed as a composition of simpler functions

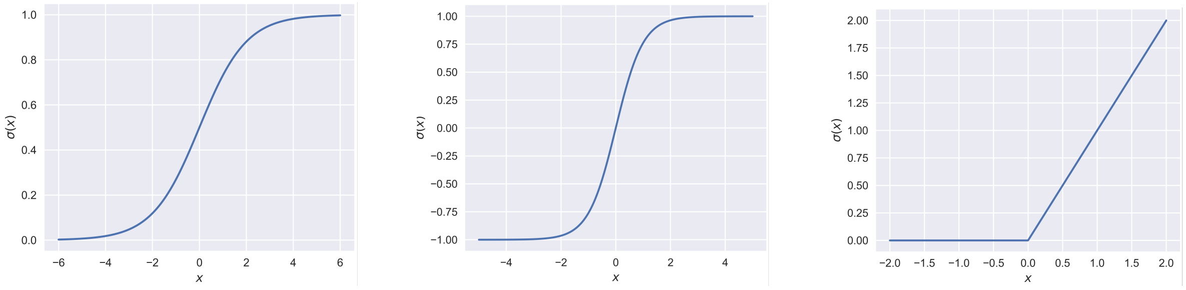
$$f(x) = \mathcal{L}_{\mathcal{D}} \circ \sigma \circ \mathcal{L}_{\mathcal{D}-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_1 \circ \sigma \circ \mathcal{L}_0(x) \;\text{ for } x \in \mathbb{R}^{p_0},$$

where $\mathcal{L}_i(x) := W_i x + b_i$ for a weight matrix $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ and bias vector $b_i \in \mathbb{R}^{p_{i+1}}$ of the i-th layer for $i = 0, 1, \dots, \mathcal{D}$. Furthermore, p_i is the width of layer i, $(p_0, p_1, \dots, p_{\mathcal{D}}, p_{\mathcal{D}+1})$ is called the *width vector* and σ is the *activation function* which is often chosen to be the rectified linear unit (ReLU) by $\sigma(x) = \max(0, x)$. The maximum is taken component-wise, that is,

$$\max(0, x) := (\max(0, x_1), \max(0, x_2), \dots, \max(0, x_n)) \;\text{ for any } x \in \mathbb{R}^n.$$

Each $\mathcal{L}_i$ in Definition 2.1 is called a *layer* of the network. The *input layer* is the first layer $\mathcal{L}_0$ and the *output layer* is the last layer $\mathcal{L}_{\mathcal{D}}$. All layers between the input and output layer are called *hidden layers*. The $\mathcal{D}$ in Definition 2.1 is called the *depth* of the network. Note that these only include the number of hidden layers and that the network has a total of $\mathcal{D} + 2$ layers. Also define $\mathcal{W} := \max\{p_1, \dots, p_{\mathcal{D}}\}$ as the *maximum width* and the *number of neurons* $\mathcal{U} := \sum_{i=1}^{\mathcal{D}} p_i$, note that the neurons in the input and output layers are not counted. Furthermore, we define the *size* of the network as the total number of parameters, that is, $\mathcal{S} := \sum_{i=0}^{\mathcal{D}} p_{i+1} \cdot (p_i + 1)$. Activation functions other than ReLU are also common. In general, an activation function is usually defined only for $x \in \mathbb{R}$, but always taken component-wise for vectors in $\mathbb{R}^n$. Commonly chosen activation functions are the sigmoid function given by $x \mapsto \frac{1}{1+e^{-x}}$, also known as the logistic function, and the hyperbolic tangent function $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$. In Fig. 2.1, these activation function are displayed. Notice how the behaviour of the hyperbolic tangent and sigmoid function are very similar. This similarity is captured in the identity $\tanh(x) = 2\sigma(2x) - 1$, where σ is the sigmoid activation function. An overview of these and more activation functions is given by Sharma et al. [2017]. In the remainder of this thesis, we will be working with the ReLU activation function.

In computer science it is common for authors to introduce neural networks in a different way than we have done. The MLP is visualized as a graph with weighted directed edges as in Fig. 2.2 and Fig. 2.3, where each edge represents a weight parameter and on each vertex, the linear combination is taken between the input of the previous layer and the weights connecting the two layers, then a bias is added and the activation function is applied. When one writes the activation of the whole layer in terms of matrix multiplication and vector additions, as is done in Fig. 2.3, one finds the same



(a) sigmoid: $x \mapsto \left(1 + e^{-x}\right)^{-1}$ (b) tanh: $x \mapsto \frac{e^{2x}-1}{e^{2x}+1}$ (c) ReLU: $x \mapsto \max(0, x)$

Figure 2.1: Three commonly used activation functions in neural networks.

expression as we have used in our definition. Thus showing that both approaches define the same structure. This approach also clarifies why the name ”neural network” is used. Notice that the width vector p in Definition 2.1 fully defines the structure of the network in Fig. 2.2, for this reason p will be also be referred to as the *architecture*.

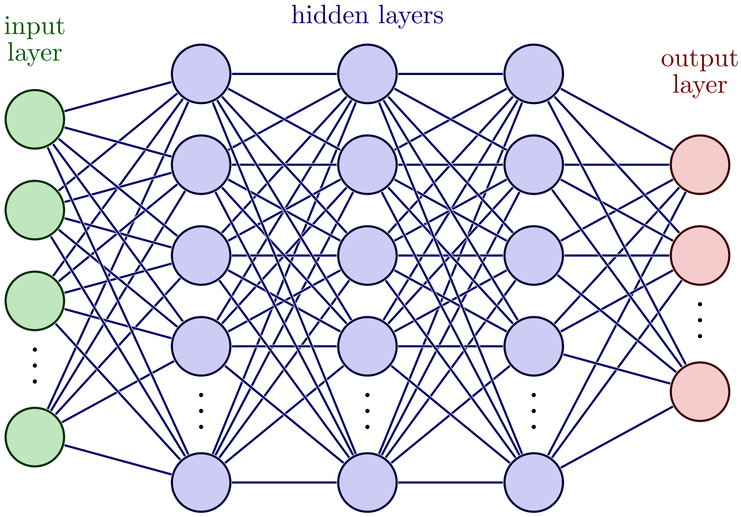
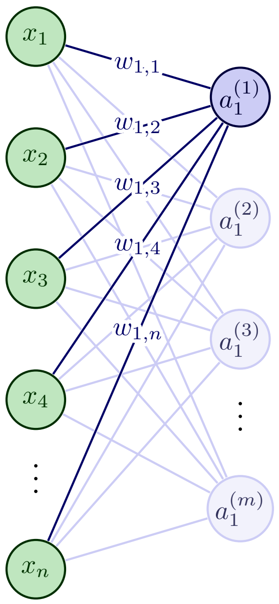


Figure 2.2: A visualization of a multi-layer perceptron [see Definition 2.1] with $\mathcal{D} = 3$. Each vertex and edge represents a neuron and weight respectively.

A useful fact about ReLU networks is that any piecewise-linear function can always be expressed using a ReLU network, where the notion of piecewise-linearity is naturally extended to real-valued functions on $\mathbb{R}^n$ [Arora et al., 2016, see Definition 3]. Conversely, any ReLU network is a piece-wise linear function. The latter follows immediately from the definition. The first statement is however less trivial, it is proven by Arora et al. [2016] in Theorem 2.1.

A network with a piecewise-linear activation function with finitely many infliction points is also a piecewise-linear function. Therefore, any network with piecewise-linear activation function can be expressed with a ReLU network. From this it follows that all error bounds and convergence results in this thesis can be generalised to networks with a piecewise-linear activation function with finitely many infliction points with suitable adjustments.



$$\begin{aligned} &= \sigma\left(w_{1,1}x_1 + w_{1,2}x_2 + \ldots + w_{1,n}x_n + b_0^{(1)}\right) \\ &= \sigma\left(\sum_{i=1}^{n} w_{1,i}x_i + b_0^{(1)}\right) \end{aligned}$$

$$\begin{pmatrix} a_1^{(1)} \\ a_1^{(2)} \\ \vdots \\ a_1^{(m)} \end{pmatrix} = \sigma\left[\begin{pmatrix} w_{1,1} & w_{1,2} & \ldots & w_{1,n} \\ w_{2,1} & w_{2,2} & \ldots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \ldots & w_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} b_0^{(1)} \\ b_0^{(2)} \\ \vdots \\ b_0^{(m)} \end{pmatrix}\right]$$

$$\begin{aligned} a^{(1)} &= \sigma \circ \mathcal{L}_0(x), \text{ where} \\ \mathcal{L}_0(x) &:= W_0 x + b_0. \end{aligned}$$

Figure 2.3: The inner workings of a neural network as defined in Definition 2.1 of the first hidden layer. The activation of the first neuron is the linear combination of the inputs x_i with the weights $w_{1,i}$, to which a bias $b_0^{(1)}$ is added and an activation function σ is applied. The activation of the whole layer can be written in terms of matrix multiplication and addition with $\mathcal{L}_0(x) = W_0x + b_0$, which aligns with the original definition of an MLP.

2.2 Network function classes.

In Definition 2.1 a width vector $p = (p_0, \ldots p_{\mathcal{D}+1})$ is used. In this thesis we focus our attention on the regression problem introduced in Section 1, that is, we try to estimate the unknown regression function $f_0 : [0,1]^d \to \mathbb{R}$ using neural networks. Hence $p_0 = d$ and $p_{\mathcal{D}+1} = 1$. From now on, we will only consider network that have input dimension d and output dimension 1. So any network architecture p with depth $\mathcal{D}$ will satisfy $p \in A_{\mathcal{D}} := \{(d, p_1, \ldots, p_{\mathcal{D}}, 1) \mid p_1, \ldots, p_{\mathcal{D}} \in \mathbb{N}\}$.

For each width vector, or architecture, there are many possible MLP's with width vector p. That is, one can choose between all possible weight matrices W_i and bias vectors b_i for $i = 1, \ldots, \mathcal{D}$. Denote the function class of all MLP's with width vector $p \in A_{\mathcal{D}}$ that are bounded by some $0 < \mathcal{B} < \infty$ by $\mathcal{NN}_p^{\mathcal{B}}$. That is,

$$\mathcal{NN}_p^{\mathcal{B}} := \{f \text{ as in Definition 2.1 } \mid f \text{ has architecture } p,\ \|f\|_\infty \le \mathcal{B}\}, \tag{2.1}$$

where $\|f\|_\infty := \sup_{x\in[0,1]^d} |f(x)|$. The architecture p is allowed to depend on the sample size n, so $p =: p_n$. This will allow the network to become deeper and wider as the sample size increases, which will give pleasant asymptotic results. Throughout this thesis, we will often omit the subscript for notational simplicity.

For a fixed depth $\mathcal{D}$, the set of all architectures with depth $\mathcal{D}$ is $A_{\mathcal{D}}$, taking the union over all depths gives the set of all possible architectures $A := \bigcup_{\mathcal{D}\ge 0} A_{\mathcal{D}}$. In order to compare different function classes $\mathcal{NN}_p^{\mathcal{B}}$ and $\mathcal{NN}_{p'}^{\mathcal{B}}$, we propose to define a partial ordering on the set of architectures A.

Definition 2.2 (Partial ordering)**.** Let X be some set. Now $(X, \le)$ is a *partially ordered* set if for all $a, b, c \in X$,

1. $a \le a$, (Reflexivity)

2. $a \leq b$ and $b \leq a \quad \Longrightarrow \quad a = b, \quad$ (Antisymmetry)

3. $a \leq b$ and $b \leq c \quad \Longrightarrow \quad a \leq c. \quad$ (Transitivity)

The relation $\leq$ is a *partial ordering* on the set X. Note that every two elements $a, b \in X$ need not be comparable. When this is the case, $\leq$ is called a *total ordering*.

Using Definition 2.2 we can define a partial ordering on the set of architectures A. Let $p, p' \in A$, assume $p \in A_{\mathcal{D}}$ and $p' \in A_{\mathcal{D}'}$. Now

$$p \leq p' \iff \mathcal{D} \leq \mathcal{D}' \text{ and } p_i \leq p'_i \text{ for all } i = 1, \dots, \mathcal{D}. \tag{2.2}$$

It is clear that this relation is indeed a partial ordering by the fact that $(\mathbb{N}, \leq)$ is a partial ordering.

Some examples of this partial ordering are $(d, 3, 1) \leq (d, 3, 1, 4, 1)$ and $(d, 1, 2, 1) \leq (d, 2, 3, 1)$. One should imagine the network with architecture p fitting inside of the network with architecture p' whenever $p \leq p'$ [see Fig. 2.2]. One might think intuitively that $\mathcal{NN}_p^{\mathcal{B}} \subseteq \mathcal{NN}_{p'}^{\mathcal{B}}$ when $p \leq p'$, since one network 'fits' inside of the other. This intuition turns out to be correct, which will be proven in Proposition 2.5. In order to prove this proposition, we shall first prove two useful lemmas.

Lemma 2.3 (Monotonicity of depth). *Consider $\mathcal{NN}_p^{\mathcal{B}}$ for some $p = (d, p_1, \dots, p_{\mathcal{D}}, 1)$ with depth $\mathcal{D} \in \mathbb{N}$. Set $p' = (d, p_1, \dots, p_{\mathcal{D}}, 1, 1)$, that is, p' has the same architecture as p but extended one layer with width 1. Now*

$$\mathcal{NN}_p^{\mathcal{B}} \subseteq \mathcal{NN}_{p'}^{\mathcal{B}},$$

that is, making the network deeper does not lose expressive power.

Proof. For any $f \in \mathcal{NN}_p^{\mathcal{B}}$, we can write f as

$$f(x) = \mathcal{L}_{\mathcal{D}} \circ \sigma \circ \mathcal{L}_{\mathcal{D}-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_1 \circ \sigma \circ \mathcal{L}_0(x) \text{ for } x \in \mathbb{R}^d.$$

Now set $a(x) := \sigma \circ \mathcal{L}_{\mathcal{D}-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_1 \circ \sigma \circ \mathcal{L}_0(x) \in \mathbb{R}^{\mathcal{D}}$ as the activation of the second to last layer of f. Observe that $a(x)_j \geq 0$ for all $j = 1, \dots, p_{\mathcal{D}}$ since $\sigma(x) = \max(0, x)$. Define $\mathcal{L}_*(x) := x = I_{\mathcal{D}} x + 0$ for any $x \in \mathbb{R}^{\mathcal{D}}$, now we can write f as

$$f(x) = \mathcal{L}_{\mathcal{D}} \circ a(x) = \mathcal{L}_{\mathcal{D}} \circ \sigma \circ \mathcal{L}_* \circ a(x).$$

We have written the function f as a network with architecture p', thus showing $f \in \mathcal{NN}_{p'}^{\mathcal{B}}$. Which proves the result. □

Lemma 2.4 (Monotonicity of width). *Fix some $j \in \{1, \dots, \mathcal{D}\}$ and consider $\mathcal{NN}_p^{\mathcal{B}}$ for some $p = (d, p_1, \dots, p_j, \dots, p_{\mathcal{D}}, 1)$ with depth $\mathcal{D} \in \mathbb{N}$. Set $p' = (d, p_1, \dots, p_j + 1, \dots, p_{\mathcal{D}}, 1)$, that is, p' has the same architecture as p but layer j is one wider than in p. Now*

$$\mathcal{NN}_p^{\mathcal{B}} \subseteq \mathcal{NN}_{p'}^{\mathcal{B}},$$

that is, making the network wider does not lose expressive power.

Proof. For any $f \in \mathcal{NN}_p^{\mathcal{B}}$, we can write f as

$$f(x) = \mathcal{L}_{\mathcal{D}} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_j \circ \sigma \circ \mathcal{L}_{j-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_0(x) \text{ for } x \in \mathbb{R}^d.$$

By making the j-th layer one wider, $\mathcal{L}_{j-1}$ and $\mathcal{L}_j$ change in structure. Write

$$\mathcal{L}_{j-1}(x) = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p_{j-1}} \\ w_{21} & w_{22} & \dots & w_{2p_{j-1}} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p_j 1} & w_{p_j 2} & \dots & w_{p_j p_{j-1}} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{p_{j-1}} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{p_j} \end{pmatrix}$$

and

$$\mathcal{L}_j(x) = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1p_j} \\ u_{21} & u_{22} & \dots & u_{2p_j} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p_{j+1}1} & u_{p_{j+1}2} & \dots & u_{p_{j+1}p_j} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{p_j} \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{p_{j+1}} \end{pmatrix}.$$

To make the j-th layer one wider without changing the function f we define

$$\mathcal{L}'_{j-1}(x) = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p_{j-1}} \\ w_{21} & w_{22} & \dots & w_{2p_{j-1}} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p_j1} & w_{p_j2} & \dots & w_{p_jp_{j-1}} \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{p_{j-1}} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{p_j} \\ 0 \end{pmatrix}$$

and

$$\mathcal{L}'_j(x) = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1p_j} & 0 \\ u_{21} & u_{22} & \dots & u_{2p_j} & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ u_{p_{j+1}1} & u_{p_{j+1}2} & \dots & u_{p_{j+1}p_j} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{p_j} \\ x_{p_j+1} \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{p_{j+1}} \end{pmatrix},$$

that is, we add zeros to the weight matrix and bias vector of $\mathcal{L}_{j-1}$ in such a way that $\mathcal{L}'_{j-1}(x) = (\text{ — } \mathcal{L}_{j-1}(x) \text{ — }, 0)^T$. Using that $\sigma(0) = 0$, it follows that

$$\begin{aligned} \mathcal{L}'_j \circ\ \sigma \circ \mathcal{L}'_{j-1}(x) &= \mathcal{L}'_j \begin{pmatrix} | \\ \sigma \circ \mathcal{L}_{j-1}(x) \\ | \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1p_j} & 0 \\ u_{21} & u_{22} & \dots & u_{2p_j} & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ u_{p_{j+1}1} & u_{p_{j+1}2} & \dots & u_{p_{j+1}p_j} & 0 \end{pmatrix} \begin{pmatrix} | \\ \sigma \circ \mathcal{L}_{j-1}(x) \\ | \\ 0 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{p_{j+1}} \end{pmatrix} = \mathcal{L}_j \circ \sigma \circ \mathcal{L}_{j-1}(x). \end{aligned}$$

By above observation we can also write f with architecture p' as

$$f(x) = \mathcal{L}_{\mathcal{D}} \circ \sigma \circ \dots \circ \sigma \circ \mathcal{L}'_j \circ \sigma \circ \mathcal{L}'_{j-1} \circ \sigma \circ \dots \circ \sigma \circ \mathcal{L}_0(x) \quad \text{for } x \in \mathbb{R}^d,$$

which proves that $f \in \mathcal{NN}^{\mathcal{B}}_{p'}$, which in turn proves the result. □

Using Lemma 2.3 and Lemma 2.4 we can prove the following proposition.

Proposition 2.5 (Monotonicity of architecture)**.** *Let $p, p' \in A$ be such that $p \leq p'$, then*

$$\mathcal{NN}^{\mathcal{B}}_p \subseteq \mathcal{NN}^{\mathcal{B}}_{p'}.$$

Proof. One should first note that Lemma 2.3 and Lemma 2.4 can be extended to an arbitrary number of extra hidden layers and wider layer respectively by induction. Since $p \leq p'$, p has at most the depth of p' and each layer of p is less wide than p'. Now it follows by the lemmas for arbitrary width and depth that $\mathcal{NN}^{\mathcal{B}}_p \subseteq \mathcal{NN}^{\mathcal{B}}_{p'}$. □

Above proposition shows that no expressive power is lost when making the architecture of the network larger in the sense of (2.2). One might be tempted to think that it is best to choose the architecture p to be as deep and wide as possible, since the network class only gains expressive power. More expressive power is however not necessarily a good thing, as it poses some practical issues.

The first problem is overfitting the data. As you make the network more complex you allow the network to have more degrees of freedom to overfit the data. Other than not making the network unnecessarily large, solutions for overfitting have been proposed like dropout [Srivastava et al., 2014] among many regularization techniques [Goodfellow et al., 2016, see ch.7].

The second problem is that training larger neural networks is more difficult. Neural networks are optimized using gradient descent methods like described in chapter 8 of Goodfellow et al. [2016]. For larger networks these optimization methods become computationally expensive and unstable, hence giving preference to smaller networks when possible.

It should be noted that Jiao et al. [2023] and Shen et al. [2021], which are both written by the same authors, use a slightly different neural network function class. We will use and build upon some of their work, so it is import to recognize the difference in the function class they define and the one defined in (2.1).

The function class $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ is the set of all MLP's that have depth $\mathcal{D}$, maximum width $\mathcal{W}$, number of neurons $\mathcal{U}$ and size $\mathcal{S}$ such that $\|f\|_\infty \leq \mathcal{B}$ for some $0 < \mathcal{B} < \infty$. Notice that the architecture of the networks in $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ is not uniquely defined since it only poses constrains on the depth, maximum width, number of neurons and size. This is a practical disadvantage without giving any benefit in return. It is therefore that we prefer to use network class $\mathcal{NN}_p^{\mathcal{B}}$ instead.

2.3 Estimation of regression function.

Remember from Section 1 that we assume

$$Y_i = f_0(X_i) + \varepsilon_i,$$

where f_0 was the unknown regression function and ε_i are i.i.d. error terms independent of X_i. This section introduces the general problem of estimating the regression function. A common approach to estimating the relation between X and Y, is to find some measurable function $f : [0,1]^d \to \mathbb{R}$ that minimizes the *loss* $L(f(X), Y)$ for some *loss function* $L : \mathbb{R}^2 \to \mathbb{R}$, note that the loss is still a random variable depending on X and Y.

Many loss functions can be chosen, a common loss function is the square loss $L(a, y) := (a - y)^2$, also known as the MSE loss or L_2 loss.

Let $Z \overset{d}{=} (X, Y)$, then for any measurable function f the *risk* $\mathcal{R}(f)$ is defined as the expected loss, that is,

$$\mathcal{R}(f) := \mathbb{E}_Z L(f(X), Y).$$

One would like to minimize this risk over the set of all measurable functions $\mathcal{M}([0,1]^d)$ to find an *optimal estimator* f^* defined by

$$f^* := \arg \min_{f \in \mathcal{M}([0,1]^d)} \mathcal{R}(f). \tag{2.3}$$

In practice, the distribution of (X, Y) is unknown and one only has access to a sample
$S = \{(X_i, Y_i)\}_{i=1}^n$ with sample size n. For any f, we define the *empirical risk* of f on the sample S as

$$\mathcal{R}_n(f) := \frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i).$$

In the context of neural networks, we would like to find the network from the function class $\mathcal{NN}_p^{\mathcal{B}}$ defined in (2.1) that minimizes this empirical risk. This estimator is called the *empirical risk minimizer*

(ERM) $\hat{f}_n$, formally defined by

$$\hat{f}_n \in \arg \min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f).$$

To evaluate the quality of any estimator f we define its *excess risk* $\mathcal{R}(f) - \mathcal{R}(f^*)$. In particular, we will show bounds on the expected excess risk $\mathbb{E}_S \left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right]$, called the *prediction error*, of the ERM $\hat{f}_n$ under various assumptions on the error ε and loss function L. In Section 2.4 we show some basic properties of the square loss function, which will be used in error bounds in Section 3 and Section 4. In Section 2.5 a general loss function is introduced that is Lipschitz continuous in both its arguments. For this Lipschitz loss we prove similar error bounds as for the square loss.

In order to provide an upper bound on the prediction error, several results from empirical process theory are used. Here we only present some basis concepts that will be needed for our upper bounds, for a more thorough overview on the theory we highly recommend van der Vaart and Wellner [1996].

We first introduce the pseudodimension, which is a measure of complexity for a function class. The following is the definition given in Bartlett et al. [2019].

Definition 2.6 (pseudodimension)**.** Let $\mathcal{F}$ be a function class of functions from $[0,1]^d$ to $\mathbb{R}$. The *pseudodimension* of $\mathcal{F}$, written $\mathrm{Pdim}(\mathcal{F})$, is the largest integer m for which there exists $x_1, \dots, x_m \in [0,1]^d$ and $y_1, \dots, y_m \in \mathbb{R}$ such that for any $b_1, \dots, b_m \in \{0,1\}$ there exists $f \in \mathcal{F}$ such that

$$f(x_i) > y_i \Leftrightarrow b_i = 1 \text{ for all } i.$$

Throughout this thesis we will work with the pseudodimension of the network class $\mathcal{NN}_p^{\mathcal{B}}$. In particular, we make use of an upper bound on the pseudodimension proven by Bartlett et al. [2019]. They showed that

$$\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}}) \leq C \cdot \mathcal{SD} \log \mathcal{S},$$

for some constant $C > 0$. Furthermore, in our upper bounds on the prediction error, we require the sample size n to be greater or equal to the pseudodimension $\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$.

Another concept from empirical process theory we use is the covering number. For any sequence $x = (x_1, \dots, x_n) \in ([0,1]^d)^n$, let

$$\mathcal{NN}_p^{\mathcal{B}}\big|_x := \{(f(x_1), \dots, f(x_n)) \big| f \in \mathcal{NN}_p^{\mathcal{B}}\}$$

denote the subset of $\mathbb{R}^n$ of evaluated points. For any positive $\delta > 0$, let $\mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\big|_x)$ denote the *covering number* of $\mathcal{NN}_p^{\mathcal{B}}\big|_x$ under the norm $\|\cdot\|_\infty$ with radius δ. That is, the smallest number of δ-balls needed to cover $\mathcal{NN}_p^{\mathcal{B}}\big|_x$. Using this, the *uniform covering number* $\mathcal{N}_n(\delta, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})$ is defined as the maximum covering number over all $x \in ([0,1]^d)^n$, that is,

$$\mathcal{N}_n(\delta, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) := \max_{x \in ([0,1]^d)^n} \mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\big|_x). \tag{2.4}$$

2.4 Least square estimation.

Under some assumptions on L, X and ε, the true regression function f_0 equals the optimal solution f^*. For instance, set $L(a,y) = (a-y)^2$ as the square loss and assume $\mathbb{E}[\varepsilon] = 0$. Then, $f^* = f_0$, which is a special case with $\mu_\varepsilon = 0$ in the following lemma.

Lemma 2.7. *Consider the regression model in (1.1) with square loss $L(a,y) = (a-y)^2$. Then, the optimal solution f^* is biased in the sense that*

$$f^* = f_0 + \mu_\varepsilon,$$

where $\mu_\varepsilon = \mathbb{E}[\varepsilon_1]$. Furthermore, $\mathcal{R}(f^*) = \mathcal{R}(f_0) - \mu_\varepsilon^2$.

Proof. Define $\tilde{f}_0 := f_0 + \mu_\varepsilon$ and $\tilde{\varepsilon}_i := \varepsilon_i - \mu_\varepsilon$ for all i. Observe that

$$Y_i = f_0(X_i) + \varepsilon_i = \tilde{f}_0(X_i) + \tilde{\varepsilon}_i$$

and $\mathbb{E}[\tilde{\varepsilon}_i] = 0$. The risk of any $f \in \mathcal{M}([0,1]^d)$ can be decomposed by independence of X and $\tilde{\varepsilon}$ as

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}\left[(f(X) - Y)^2\right] = \mathbb{E}\left[(f(X) - \tilde{f}_0(X) - \tilde{\varepsilon})^2\right] \\ &= \mathbb{E}\left[(f(X) - \tilde{f}_0(X))^2\right] + \mathbb{E}[\tilde{\varepsilon}]\mathbb{E}\left[f(X) - \tilde{f}_0(X)\right] + \mathbb{E}\left[\tilde{\varepsilon}^2\right] = \mathbb{E}\left[(f(X) - \tilde{f}_0(X))^2\right] + \mathbb{E}\left[\tilde{\varepsilon}^2\right].\end{aligned}$$

From this it is concluded that $f^* = \tilde{f}_0 = f_0 + \mu_\varepsilon$ since $\mathcal{R}(\tilde{f}_0) = \mathbb{E}\left[\tilde{\varepsilon}^2\right]$ and $\mathcal{R}(f) \geq \mathbb{E}\left[\tilde{\varepsilon}^2\right]$ for all $f \in \mathcal{M}([0,1]^d)$.

The final statement follows by direct calculation,

$$\mathcal{R}(f^*) = \mathbb{E}\left[(f^*(X) - Y)^2\right] = \mathbb{E}\left[([f_0(X) - Y] + \mu_\varepsilon)^2\right] = \mathcal{R}(f_0) - \mu_\varepsilon^2.$$

□

Lemma 2.7 shows that the square loss is an excellent choice when one is certain that the noise ε has zero mean, since a consistent estimator for f^* will be consistent with respect to the unknown regression function f_0. At the same time, above lemma shows that the whole estimation approach is flawed from the beginning in the sense that estimating f^* will result in a biased estimator if $\mu_\varepsilon \neq 0$. It begs the question whether a better loss function exists such that f^* is close to f_0 even if the noise has non-zero mean. This question remains unanswered throughout this thesis, though it is definitely worth further exploration.

Lemma 2.8. *Consider the regression model in (1.1) and denote $L(a,y) = (a-y)^2$ as the square loss. For any random sample $S = \{(X_i, Y_i)\}_{i=1}^n$ and function $f : [0,1]^d \to \mathbb{R}$, depending possibly on the random sample S,*

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \|f - f^*\|_{L^2(\nu)}^2 := \mathbb{E}_X\left[(f(X) - f^*(X))^2\right],$$

where ν denotes the density of X. Furthermore, if one assumes $\mathbb{E}[\varepsilon_1] = 0$, the prediction error of the ERM $\hat{f}_n \in argmin_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ satisfies

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f_0)\right] &= \mathbb{E}_S\left[\|\hat{f}_n - f_0\|_{L^2(\nu)}^2\right] \\ &\leq \mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] + 2 \inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|\ f - f_0\|_{L^2(\nu)}^2.\end{aligned}$$

Proof. The proof is originally given by Jiao et al. [2023] in Lemma 3.1 only for $\mathbb{E}[\varepsilon_1] = 0$. We present the proof with more details to accompany the reader, we also generalize the first statement to $\mathbb{E}[\varepsilon_1] \neq 0$.

First we consider the case when $\mathbb{E}[\varepsilon_1] = 0$. By direct calculation it holds that for any f, depending possibly on the random sample S,

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f_0) &= \mathbb{E}_Z[L(f(X), Y) - L(f_0(X), Y)] \\ &= \mathbb{E}_Z\left[f(X)^2 - 2f(X)Y + Y^2 - f_0(X)^2 + 2f_0(X)Y - Y^2\right] \\ &= \mathbb{E}_Z\left[f(X)^2 - 2f(X)Y - f_0(X)^2 + 2f_0(X)Y\right]\end{aligned}$$

Observe that $\mathbb{E}_Z\left[Yf_0(X)\right] = \mathbb{E}_Z\left[f_0(X)^2\right] + \mathbb{E}_Z\left[\varepsilon f_0(X)\right] = \mathbb{E}_Z\left[f_0(X)^2\right]$ since ε is independent of X. Similarly, $\mathbb{E}_Z\left[Yf(X)\right] = \mathbb{E}_Z\left[f_0(X)f(X)\right]$, which implies

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f_0) &= \mathbb{E}_X\left[f(X)^2 - 2f(X)f_0(X) + f_0(X)^2\right] = \mathbb{E}_X\left[(f(X) - f_0(X))^2\right] \\ &= \|f - f_0\|_{L^2(\nu)}^2. \qquad (2.5)\end{aligned}$$

Now assume $\mu_\varepsilon := \mathbb{E}[\varepsilon_1] \neq 0$, set $\tilde{f}_0 = f_0 + \mu_\varepsilon$ and $\tilde{\varepsilon}_i = \varepsilon_i - \mu_\varepsilon$ for any i. Observe that $\mathbb{E}[\tilde{\varepsilon}_i] = 0$ and

$$Y_i = f_0(X_i) + \varepsilon_i = \tilde{f}_0(X_i) + \tilde{\varepsilon}_i.$$

Hence by the just proven equality it follows that $\mathcal{R}(f) - \mathcal{R}(\tilde{f}_0) = \|f - \tilde{f}_0\|^2_{L^2(\nu)}$. By Lemma 2.7 we have $\tilde{f}_0 = f_0 + \mu_\varepsilon = f^*$, which proves the first result.

By definition of the ERM, $\mathcal{R}_n(\hat{f}_n) \leq \mathcal{R}_n(f)$ for any $f \in \mathcal{NN}_p^{\mathcal{B}}$. Hence

$$\mathcal{R}_n(\hat{f}_n) - \mathcal{R}_n(f_0) \leq \mathcal{R}_n(\bar{f}) - \mathcal{R}_n(f_0),$$

where $\bar{f} \in \arg\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f_0\|^2_{L^2(\nu)}$. Taking expectation on both sides we get

$$\begin{aligned}
\mathbb{E}_S\left[\mathcal{R}_n(\hat{f}_n) - \mathcal{R}(f_0)\right] &\leq \mathcal{R}(\bar{f}) - \mathcal{R}(f_0) = \mathbb{E}_Z\left[Y^2 - 2\bar{f}(X)Y + \bar{f}(X)^2 - Y^2 + 2f_0(X)Y - f_0(X)^2\right] \\
&= \mathbb{E}_Z\left[-2\bar{f}(X)f_0(X) + \bar{f}(X)^2 + f_0(X)^2\right] \\
&= \mathbb{E}_X\left[(\bar{f}(X) - f_0(X))^2\right] = \|\bar{f} - f_0\|^2_{L^2(\nu)}.
\end{aligned}$$

Since $\bar{f} \in \arg\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f_n - f_0\|^2_{L^2(\nu)}$, we thus have

$$\mathbb{E}_S\left[\mathcal{R}_n(\hat{f}_n) - \mathcal{R}(f_0)\right] \leq \inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f_0\|^2_{L^2(\nu)}. \tag{2.6}$$

Multiplying (2.6) by 2 and adding (2.5) we obtain

$$\mathbb{E}_S\left[\|\hat{f}_n - f_0\|^2_{L^2(\nu)}\right] + 2\mathbb{E}_S\left[\mathcal{R}_n(\hat{f}_n) - \mathcal{R}(f_0)\right] \leq \mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f_0)\right] + 2\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f_0\|^2_{L^2(\nu)}.$$

Rearranging terms gives

$$\begin{aligned}
\mathbb{E}_S\left[\|\hat{f}_n - f_0\|^2_{L^2(\nu)}\right] &\leq \mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + 2\mathcal{R}(f_0)\right] + 2\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f_0\|^2_{L^2(\nu)} \\
&= \mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] + 2\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f_0\|^2_{L^2(\nu)},
\end{aligned}$$

which proves the last inequality. □

Lemma 2.8 shows that the excess risk is not just the difference in risk, it is in fact a distance in the sense of a metric [Munkres, 2000, see section 20]. Showing that the excess risk of any estimator f converging to zero, is equivalent to showing convergence of f to f^* in $L^2(\nu)$, which is a much stronger result.

2.5 Lipschitz loss estimation.

Lemma 2.7 shows that under the square loss, the optimal estimator f^* is biased in the sense that $f^* = f_0 + \mu_\varepsilon$. It would be better if f^* is equal, or at least close to, f_0. To achieve this one can consider a different loss function, a *robust* loss function $L : \mathbb{R}^2 \to \mathbb{R}$. Shen et al. [2021] give several examples of robust loss functions.

- Least absolute deviation (LAD) loss: $L(a, y) = |a - y|,\ (a, y) \in \mathbb{R}^2$.
- Quantile loss: $L(a, y) = \rho_\tau(a - y),\ (a, y) \in \mathbb{R}^2$, where

$$\rho_\tau(x) = \begin{cases} \tau x & \text{if } x \geq 0 \\ (\tau - 1)x & \text{if } x < 0 \end{cases} \quad \text{for some } \tau \in (0, 1).$$

- Huber loss: $L(a,y) = h_\zeta(a-y),\ (a,y)\in\mathbb{R}^2$, where

$$h_\zeta(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \le \zeta \\ \zeta|x| - \frac{\zeta^2}{2} & \text{if } |x| > \zeta \end{cases} \quad \text{for some } \zeta > 0.$$

- Cauchy loss: $L(a,y) = \log\left(1+\kappa^2(a-y)^2\right),\ (a,y)\in\mathbb{R}^2$, for some $\kappa > 0$.
- Tukey's biweight loss: $L(a,y) = T_t(a-y),\ (a,y)\in\mathbb{R}^2$, where

$$T_t(x) = \begin{cases} \frac{t^2}{6}\left[1-\left\{1-\left(\frac{x}{t}\right)^2\right\}^3\right] & \text{if } |x| \le t \\ \frac{t^2}{6} & \text{if } |x| > t \end{cases} \quad \text{for some } t > 0.$$

All above loss function can be written as a function of the difference $a-y$, Fig. 2.4 displays them as a function of this difference. Furthermore, all loss functions are continuous and λ_L-Lipschitz in both arguments, that is,

$$\begin{aligned} \left|L(a_1,\cdot) - L(a_2,\cdot)\right| &\le \lambda_L\left|a_1-a_2\right|, \\ \left|L(\cdot,y_1) - L(\cdot,y_2)\right| &\le \lambda_L\left|y_1-y_2\right|, \end{aligned}$$

for any $a_1,a_2,y_1,y_2\in\mathbb{R}$.

One would like to have an identity relating f^* and f_0 as is given in Lemma 2.7 for the square loss. For this general Lipschitz continuous, we cannot formulate such a relation. However, the following does hold.

Lemma 2.9. *Consider the regression model in (1.1) with some continuous loss function L. Assume that L can be written as a difference between its two inputs, that is, $L(a,y) = \psi(a-y)$ for some function ψ. Also assume that this function ψ is symmetric, differentiable, monotonically increasing on $[0,\infty)$, and strictly monotonically increasing in at least some neighbourhood. Then, if $\mathbb{E}[\varepsilon] < \infty$, and ε has a symmetric density that is decreasing on $[0,\infty)$, the optimal solution f^* satisfies*

$$f^* = f_0.$$

Indeed, the loss functions we have discussed above, except for the quantile loss, satisfy the assumptions of above Lemma 2.9. Hence, if ε is symmetric with existing mean, the optimal solution is robust against possible symmetric outliers, like those generated by a t-distribution.

Proof. For any f, using the tower property, the risk of can be written as

$$\begin{aligned} \mathcal{R}(f) &= \mathbb{E}_Z\left[L(f(X),Y)\right] = \mathbb{E}_Z\left[\psi(f(X)-Y)\right] = \mathbb{E}_{(X,\varepsilon)}\left[\psi(f(X)-f_0(X)-\varepsilon)\right] \\ &= \mathbb{E}_{(X,\varepsilon)}\left[\psi(\varepsilon-\{f(X)-f_0(X)\})\right] = \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left(\psi(\varepsilon-\{f(X)-f_0(X)\})\,\middle|\,X\right)\right]. \end{aligned}$$

Instead of minimizing this expression directly, we consider the function $\lambda(\mu) := \mathbb{E}_\varepsilon\psi(\varepsilon-\mu)$ for any $\mu\in\mathbb{R}$. In the following, it is proven that λ has a unique minimum at $\mu = 0$. This statement was originally proven in Maronna et al. [2019, Thm 10.2].

Observe that

$$\lambda'(\mu) = -\int_{\mathbb{R}} f_\varepsilon(x)\psi'(x-\mu)dx.$$

While doing some substitutions, using the fact that ψ' is odd and f_ε even, this can be written as

$$\begin{aligned}
\lambda'(\mu) &= -\int_{\mu}^{\infty} f_\varepsilon(x)\psi'(x-\mu)dx - \int_{-\infty}^{\mu} f_\varepsilon(x)\psi'(x-u)dx \\
&= -\int_{0}^{\infty} f_\varepsilon(x+\mu)\psi'(x)dx + \int_{-\infty}^{\mu} f_\varepsilon(x)\psi'(\mu-x)dx \\
&= -\int_{0}^{\infty} f_\varepsilon(x+\mu)\psi'(x)dx + \int_{0}^{\infty} f_\varepsilon(x-\mu)\psi'(x)dx \\
&= \int_{0}^{\infty} \psi'(x)\left[f_\varepsilon(x-\mu) - f_\varepsilon(x+\mu)\right]dx.
\end{aligned}$$

Observe from this that $\lambda'(-\mu) = -\lambda'(\mu)$ for all μ and $\lambda'(0) = 0$. We now show that $\lambda'(\mu) > 0$ for $\mu > 0$.

If x and μ are positive, then $|x-\mu| < |x+\mu|$, hence it follows that $f_\varepsilon(x-\mu) > f_\varepsilon(x+\mu)$ by assumption on f_ε. Also note that by assumption on ψ, we have $\psi'(x) \geq 0$ for all $x \geq 0$ and $\psi'(x) > 0$ for all $x \in (a,b)$ for some $0 \geq a < b$. Since ψ' is strictly positive in at least some positive neighbourhood, and $f_\varepsilon(x-\mu) > f_\varepsilon(x+\mu)$, we conclude that the integral must be positive. That is, $\lambda'(\mu) > 0$ for $\mu > 0$. By being odd, $\lambda'(\mu) < 0$ for $\mu < 0$. Hence the only minimum is $\mu = 0$.

Above segment shows that for any $X \in [0,1]^d$, $\mathbb{E}_\varepsilon\left(\psi(\varepsilon - \{f(X) - f_0(X)\}) \,\middle|\, X\right)$ has a unique minimum at $f^*(X) = f_0(X)$. Thus,

$$\mathbb{E}_\varepsilon\left(\psi(\varepsilon - \{f^*(X) - f_0(X)\}) \,\middle|\, X\right) \leq \mathbb{E}_\varepsilon\left(\psi(\varepsilon - \{f(X) - f_0(X)\}) \,\middle|\, X\right),$$

for all $X \in [0,1]^d$ and all measurable functions f. Note that if $f \leq g$ for two positive and measurable functions f and g, one has $\mathbb{E}[f(X)] \leq \mathbb{E}[g(X)]$. From this it follows that $f^* = f_0$ is also the unique minimum of the risk $\mathcal{R}$ over all measurable function f. □

In Section 5 any loss function can be considered, as long as it is λ_L-Lipschitz in both its arguments, and $L(a,y) = 0$ for any $a = y \in \mathbb{R}$. All just mentioned loss functions satisfy these properties. Furthermore, Shen et al. [2021] give the value of the Lipschitz constant λ_L, along with information about continuity, convexity and differentiability in Table 2. Without any further proof, we will use the information from Table 2 throughout Section 5 and Section 6.

Source: Shen et al. [2021]

	LAD	Quantile	Huber	Cauchy	Tukey
Hyper parameter	NA	$\tau \in (0,1)$	$\zeta > 0$	$\kappa > 0$	$t > 0$
λ_L	1	$\max(\tau, 1-\tau)$	ζ	κ	$\frac{16t}{25}\sqrt{5}$
Continuous	TRUE	TRUE	TRUE	TRUE	TRUE
Convex	TRUE	TRUE	TRUE	FALSE	FALSE
Differentiable	FALSE	FALSE	TRUE	TRUE	TRUE

Table 2: An overview of different robust loss functions. Note that ”NA” stands for not applicable.

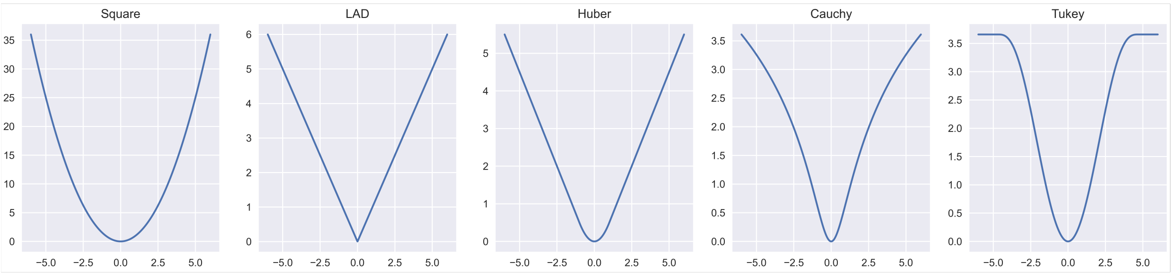


Figure 2.4: Different loss functions that are λ_L-Lipschitz in both its arguments and satisfy $L(x,x) = 0$ for any $x \in \mathbb{R}$, along with the square loss, which is not Lipschitz continuous. For the shown Huber loss, $\zeta = 1$; Cauchy loss, $\kappa = 1$; Tukey loss, $t = 4.685$. The chosen value for t is copied from Belagiannis et al. [2015]. Each loss function can be expressed a function of the difference $a - y$, for any $y, a \in \mathbb{R}$. The loss functions are displayed as a function of this difference.

2.6 Modelling outliers.

To model the outliers in the regression model from (1.1), we will first have to introduce so called mixture distributions.

Consider an event A with $\mathbb{P}(A) =: \alpha$, where an outlier is observed, that is, ε takes an extreme value. To model these large values we use some outlier distribution function F, meaning ε takes values with distribution function F on event A. Conversely, on the complement A^c a normal observation is made. A commonly used distribution is a normal distribution with mean zero and finite variance σ^2, denoted by Φ_σ, where ε takes values with distribution function Φ_σ on A^c. Now by the total law of probability

$$F_\varepsilon(t) = \mathbb{P}(\varepsilon \leq t) = \mathbb{P}(\varepsilon \leq t | A^c)\mathbb{P}(A^c) + \mathbb{P}(\varepsilon \leq t | A)\mathbb{P}(A) = \Phi_\sigma(t)(1-\alpha) + F(t)\alpha.$$

Such a combination of distributions is called a *mixture distribution.* Note that, if F is differentiable, ε has density

$$f_\varepsilon(t) = \frac{dF_\varepsilon(t)}{dt} = (1-\alpha)f(t) + \alpha\varphi_\sigma(t),$$

where f is the density of the outlier distribution and φ_σ is the density of Φ_σ. The contamination model just described is known as the *Tukey-Huber* model [Maronna et al., 2019, see p.19].

Various choices for outliers distributions can be made. A normal distribution with higher variance, say $10\sigma^2$, could be used. To get even more extreme outliers a t-distribution or Fréchet distribution can be used, where the Fréchet has non zero mean since it is only defined for positive values. The theoretical results will be formulated without specifying an outlier distribution, but in the simulations [see Section 6], the three just described distributions will be used. Table 3 gives a brief overview of relevant properties of these distributions.

		expectation	moments finite
normal distribution	$\mathrm{N}(0, 10\sigma^2)$	0	all moments are finite
t-distribution	$t(\nu)$	0 if $\nu > 1$	k-th moment finite for $k < \nu$
Fréchet distribution	Fréchet(λ)	$\Gamma\left(1 - \frac{1}{\lambda}\right)$ if $\lambda > 1$	k-th moment finite for $k < \lambda$

Table 3: An overview of possible outlier densities with some basic properties.

3 Square loss with sub-exponential error.

In this section we present an error bound on the prediction error $\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right]$ under the assumption that ε is sub-exponential. Furthermore, we show that the ERM $\hat{f}_n$ is a consistent estimator for the optimal estimator f^*. Before the prediction error bound is covered, sub-exponential random variables are introduced together with their properties.

3.1 Sub-exponential distributions and their properties.

Let us start with defining when a random variable is sub-exponential.

Definition 3.1 (Sub-exponential random variable)**.** A real-valued random variable X is *sub-exponential* if there exists a $K > 0$ such that

$$\mathbb{E}\exp(\lambda|X|) \leq \exp(K\lambda) \quad \text{for all } \lambda \text{ with } 0 \leq \lambda \leq \frac{1}{K}.$$

It should be emphasized that this definition does not define the distribution of X, it is only a property that X can have.

The given definition does not give much insight into the behaviour of sub-exponential random variables. However, the following proposition from Vershynin [2018, page 32] gives a few equivalent properties for sub-exponential random variables.

Proposition 3.2 (Sub-exponential properties)**.** *Let X be a random variable. If any of the following conditions holds, then X is sub-exponential. Furthermore, all conditions are equivalent.*

(i) There exists a $K_1 > 0$ such that the tails of X satisfy

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-tK_1) \textit{ for all } t \geq 0.$$

(ii) There exists a $K_2 > 0$ such that the moments of X satisfy

$$\|X\|_{L^p} := (\mathbb{E}|X|^p)^{\frac{1}{p}} \leq K_2 p \textit{ for all } p \geq 1.$$

(iii) There exists a $K_3 > 0$ such that the moment generating function of $|X|$ satisfies

$$\mathbb{E}\exp(\lambda|X|) \leq \exp(K_3\lambda) \quad \textit{for all } \lambda \textit{ with } 0 \leq \lambda \leq \frac{1}{K_3}.$$

(iv) There exists a $K_4 > 0$ such that the moment generating function of $|X|$ is bounded at some point, namely

$$\mathbb{E}\exp(K_4|X|) \leq 2.$$

Moreover, if $\mathbb{E}X = 0$ then properties $(i) - (iv)$ are also equivalent to the following one.

(v) There exists a $K_5 > 0$ such that the moment generating function of $X - \mathbb{E}X$ satisfies

$$\mathbb{E}\exp(\lambda X) \leq e^{K_5^2\lambda^2} \quad \textit{for all } \lambda \textit{ with } |\lambda| \leq \frac{1}{K_5}.$$

Proof. The proof is given by Vershynin [2018, page 32]. □

Remark. Property (v) shows that X has all moments finite, since the moment generating function $M_X(t) := \mathbb{E}[e^{tX}]$ exists in a neighbourhood around zero and

$$\mathbb{E}[X^n] = \left.\frac{d^n M_X}{dt^n}\right|_{t=0}.$$

Property (i) shows that the tails of X must decay exponentially, hence the name "sub-exponential". This observation is the original motivation for the definition and demonstrates that normal and exponential distributions are sub-exponential, given that their tails decay with rates $\exp(-t^2)$ and $\exp(-t)$ respectively. We will delve into more detailed examples later. First we derive some valuable properties of sub-exponential random variables, namely that the set of sub-exponential random variables is closed under finite linear combinations.

Proposition 3.3. *Let $a_1, \dots, a_n \in \mathbb{R}$ and $X_1, \dots X_n$ be random variables such that every X_i is sub-exponential. Then $\sum_{i=1}^n a_i X_i$ is also sub-exponential.*

Proof. We first prove that for any sub-exponential random variable X, it follows that aX is sub-exponential for any $a \in \mathbb{R}$. By property (iv) of Proposition 3.2 there exists a $K > 0$ such that

$$\mathbb{E}\exp(K|X|) \leq 2.$$

Set $\tilde{K} := \frac{K}{|a|} > 0$. Then

$$\mathbb{E}\exp(\tilde{K}|aX|) = \mathbb{E}\exp(\tilde{K}|a||X|) = \mathbb{E}\exp(K|X|) \leq 2.$$

Thus by property (iv) of Proposition 3.2 we conclude that aX is sub-exponential. Now let X_1 and X_2 be two sub-exponential random variables and define $\alpha := \frac{1}{2}\min(\alpha_1, \alpha_2)$ where $\alpha_1, \alpha_2 > 0$ are such that $\mathbb{E}\exp(\alpha_1|X_1|) \leq 2$ and $\mathbb{E}\exp(\alpha_2|X_2|) \leq 2$ by property (iv) of Proposition 3.2. By using Cauchy-Schwarz it follows that

$$\begin{aligned}\mathbb{E}\exp(\alpha|X_1 + X_2|) &\leq \mathbb{E}[\exp(\alpha|X_1|)\exp(\alpha|X_2|)] \leq \sqrt{\mathbb{E}[\exp(2\alpha|X_1|)]\mathbb{E}[\exp(2\alpha|X_2|)]} \\ &\leq \sqrt{\mathbb{E}[\exp(\alpha_1|X_1|)]\mathbb{E}[\exp(\alpha_2|X_2|)]} \leq \sqrt{2 \cdot 2} = 2,\end{aligned}$$

hence $X_1 + X_2$ is sub-exponential. Combining the two results above and using induction we conclude that any finite linear combination of sub-exponential random variables is sub-exponential. □

Corollary 3.3.1. *Let X be a sub-exponential random variable and $C \in \mathbb{R}$. Then $X + C$ is again sub-exponential.*

Proof. Note that $C = \delta_C$ where δ_C is the point mass at C. Then for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[e^{\lambda|\delta_C|}] = e^{\lambda|C|}.$$

Hence by property (iii) of Proposition 3.2 it follows that δ_C is sub-exponential. By applying Proposition 3.3 to $X + C = X + \delta_C$ the result follows. □

Note that Corollary 3.3.1 shows that X is sub-exponential if and only if $X - \mathbb{E}X$ is sub-exponential. This can be a computational benefit when showing a random variable is sub-exponential. The result is also heavily used in the proof of Corollary 3.6.1.

Example. Let $X \sim N(\mu, \sigma^2)$ and $Y := X - \mu$, then $Y \sim N(0, \sigma^2)$. Now

$$\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) = \mathbb{E}\exp(\lambda Y) = \exp\left(\frac{\sigma^2}{2}\lambda^2\right) \text{ for all } \lambda \in \mathbb{R}.$$

Hence by property (v) of Proposition 3.2 any normal distribution is sub-exponential.

Example. Let $X \sim \text{Exp}(\lambda)$. Then

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(X \geq t) = e^{-\lambda t} \leq 2e^{-t\lambda} \text{ for all } \lambda \in \mathbb{R}.$$

Thus by property (i) of Proposition 3.2 we conclude that X is sub-exponential.

Example. Let $X_1, \dots X_n \sim \mathrm{Exp}(\lambda)$ be n i.i.d. random variables, then $\sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$ [see Grimmett and Welsh, 2014, ex. 10, p. 103]. By previous example X_i is sub-exponential for all i, hence $\sum_{i=1}^n X_i$ is also sub-exponential by Proposition 3.3. We conclude that $\Gamma(n, \lambda)$ is sub-exponential for any $n \in \mathbb{N}$ and $\lambda > 0$.

With the preceding examples we have demonstrated that many common distribution are sub-exponential. However, there remains a big class of sub-exponential distributions that we have not covered, the set of bounded distributions. Let X be a random variable that is bounded. X does not have tails, hence the tails decay exponentially. Therefore we expect X to be sub-exponential. To prove this we first introduce the notion of sub-Gaussian random variables and Hoeffding's inequality.

Definition 3.4 (Sub-Gaussian random variable)**.** A real-valued random variable X is *sub-Gaussian* if there exists a $K > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2K^2) \quad \text{for all } t \geq 0.$$

A similar proposition to Proposition 3.2 exists for sub-Gaussian random variables. For more information on sub-Gaussian random variables it is recommended to read chapter 2.5 from Vershynin [2018]. For our purposes the following lemma [Vershynin, 2018, p. 34] is sufficient.

Lemma 3.5 (Sub-exponential is sub-Gaussian squared)**.** *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential.*

Example. Let $X_1 \dots X_n$ be n i.i.d. standard normal random variables. The tail of X_i can be bounded by

$$\begin{aligned}
\mathbb{P}(|X_i| \geq t) = 2\mathbb{P}(X_i \geq t) &= 2\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx. \text{ Set } u := x - t, \text{ then} \\
&= 2\int_{u=0}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(u+t)^2}{2}\right) du \leq 2\int_{u=0}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} e^{-\frac{1}{2}u^2} du \\
&= 2e^{-\frac{1}{2}t^2} \int_{u=0}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = 2e^{-\frac{1}{2}t^2}\mathbb{P}(X_i \geq 0) \leq 2e^{-\frac{1}{2}t^2} \quad \text{for all } t \geq 0.
\end{aligned}$$

So X_i is sub-Gaussian by definition for all i. From Lemma 3.5 and Proposition 3.3 we know that $Z := \sum_{i=1}^n X_i^2$ is sub-exponential, where Z has by definition a χ^2 distribution with n degrees of freedom. Thus we conclude that any χ^2 distribution is sub-exponential.

Theorem 3.6 (Hoeffding's inequality)**.** *Let $X_1, \dots X_n$ be independent random variables. Assume that $X_i \in [a_i, b_i]$ a.s. for all i. Then for any $t \geq 0$ we have*

$$\mathbb{P}\left[|\sum_{i=1}^n (X_i - \mathbb{E}X_i)| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Corollary 3.6.1. *If X is a random variable such that $X \in [a, b]$ then X is sub-exponential.*

Proof. Define $Y := X - a$ and observe $Y \in [0, c]$ where $c := b - a$. By Hoeffding's inequality we have

$$\mathbb{P}[|\sqrt{Y}| \geq t] \leq 2\exp\left(-\frac{2t^2}{(\sqrt{c})^2}\right) = 2\exp\left(-\frac{2}{c}t^2\right).$$

This implies that $\sqrt{Y}$ is sub-Gaussian by definition, which is equivalent to Y being sub-exponential by Lemma 3.5. Since

$$X = Y + a,$$

we conclude by Corollary 3.3.1 that X is sub-exponential. □

Thus far we have seen that sub-exponential random variables have nice properties. Furthermore, many common distribution such as normal, chi-squared, exponential, gamma and bounded distributions are all sub-exponential. As mentioned before, being sub-exponential implies that all moments are finite. Therefore any distribution that does not have all moments finite, is not sub-exponential. Examples are Cauchy, Pareto and t-distributions.

3.2 Prediction error bound for sub-exponential error and square loss.

In the following we derive an upper bound on the prediction error $\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right]$, we then show the empirical risk minimizer is consistent with respect to the optimal estimator f^*. The results in this subsection hold for sub-exponential error ε and are originally proved by Jiao et al. [2023], though we present a different proof and a slightly different statement in Theorem 3.8 than the theorem given in Jiao et al. [2023].

Lemma 3.7. *Consider the regression model from (1.1). Assume that ε is sub-exponential and $\|f_0\|_\infty \leq \mathcal{B} - |\mu_\varepsilon|$ with $\mathcal{B} \geq 1$ and $\mu_\varepsilon := \mathbb{E}[\varepsilon_1]$. Let $\hat{f}_n \in \arg\min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ denote the empirical risk minimizer over $\mathcal{NN}_p^{\mathcal{B}}$ and $L(a, y) = (a - y)^2$ the square loss. Then, for $n \geq \frac{1}{2}\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$,*

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right] &= \mathbb{E}_S\|\hat{f}_n - f^*\|^2_{L^2(\nu)} \\ &\leq C_0\mathcal{B}^5(\log n)^5\frac{1}{n}\mathcal{S}\mathcal{D}\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\{\mathcal{R}(f) - \mathcal{R}(f^*)\} \\ &= C_0\mathcal{B}^5(\log n)^5\frac{1}{n}\mathcal{S}\mathcal{D}\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\|f - f^*\|^2_{L^2(\nu)},\end{aligned}$$

where C_0 is a constant independent of $d, n, \mathcal{B}, \mathcal{D}, \mathcal{W}$ and $\mathcal{S}$.

Proof. We first consider the case when $\mathbb{E}[\varepsilon] = 0$. By Lemma 3.2 from Jiao et al. [2023], we have, for all $n \geq \frac{1}{2}\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$,

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq C_0\mathcal{B}^5(\log n)^5\frac{1}{n}\mathcal{S}\mathcal{D}\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\|f - f_0\|^2_{L^2(\nu)}.$$

By Lemma 2.7 we have $f_0 = f^*$ and by Lemma 2.8 we have

$$\|f - f_0\|^2_{L^2(\nu)} = \mathcal{R}(f) - \mathcal{R}(f_0)$$

for any f, which proves the result for $\mathbb{E}[\varepsilon] = 0$.

Whenever $\mathbb{E}[\varepsilon] \neq 0$, set $\tilde{f}_0 := f_0 + \mu_\varepsilon$ and $\tilde{\varepsilon}_i := \varepsilon_i - \mu_\varepsilon$. Observe that

$$Y_i = f_0(X_i) + \varepsilon_i = \tilde{f}_0(X_i) + \tilde{\varepsilon}_i \text{ for all } i$$

and $\mathbb{E}[\tilde{\varepsilon}] = 0$. Also note that

$$\|\tilde{f}_0\|_\infty = \|f_0 + \mu_\varepsilon\|_\infty \leq \mathcal{B} + |\mu_\varepsilon| - |\mu_\varepsilon| = \mathcal{B},$$

hence by the just proven result for zero-mean error, we obtain the result since $\tilde{f}_0 = f^*$ by Lemma 2.7. □

Lemma 3.7 gives a non-asymptotic upper bound on the prediction error $\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right]$. The upper bound consists of two terms. The second term is clearly zero if f^* can be expressed directly as an MLP in the function class $\mathcal{NN}_p^{\mathcal{B}}$. Note that the first term converges to 0 as $n \to \infty$ if the architecture of p is fixed. One should observe that taking the architecture fixed also fixes the second term. Hence a trade-off must be made between making sure the first term converges to zero with the sample size, and allowing the network to grow with the sample size to decrease the second term. The following theorem shows that both terms can converge to zero with the sample size under suitable assumptions.

Theorem 3.8 (consistency of ERM under square loss and sub-exponential error)**.** *Consider the regression model in (1.1). Assume that ε is sub-exponential, f_0 is continuous on $[0,1]^d$, $\|f_0\|_\infty \leq \mathcal{B} - |\mu_\varepsilon|$ with $\mathcal{B} \geq 1$ and $\mu_\varepsilon := \mathbb{E}[\varepsilon_1]$. Also assume that each layer of p has at least width $d+1$ and p has depth $\mathcal{D} \geq 3$. Let $\hat{f}_n \in \arg\min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ denote the empirical risk minimizer over $\mathcal{NN}_p^{\mathcal{B}}$ and $L(a,y) = (a-y)^2$ the square loss. If the architecture p satisfies,*

$$\mathcal{S} \to \infty \quad \textit{and} \quad \mathcal{B}^5 (\log n)^5 \frac{1}{n} \mathcal{S}\mathcal{D} \log \mathcal{S} \to 0 \quad \textit{as} \quad n \to \infty.$$

Then, the prediction error of the ERM $\hat{f}_n$ satisfies

$$\lim_{n\to\infty} \mathbb{E}_S \left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right] = \lim_{n\to\infty} \mathbb{E}_S \left[\|\hat{f}_n - f^*\|_{L^2(\nu)}^2 \right] = 0.$$

Furthermore, if $f_0 \in \mathcal{NN}_p^{\mathcal{B}}$, the condition that $\mathcal{S} \to \infty$ can be dropped.

Proof. First observe that for n large enough, we have

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right] \leq C_0 \mathcal{B}^5 (\log n)^5 \frac{1}{n} \mathcal{S}\mathcal{D}\log\mathcal{S} + 2 \inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \left\{ \mathcal{R}(f) - \mathcal{R}(f^*) \right\},$$

by Lemma 3.7. Hence it suffices to show that

$$\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \left\{ \mathcal{R}(f) - \mathcal{R}(f^*) \right\} = \inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f^*\|_{L^2(\nu)}^2 \to 0 \quad \text{as} \quad n \to \infty,$$

since $\mathcal{B}^5 (\log n)^5 \frac{1}{n} \mathcal{S}\mathcal{D}\log\mathcal{S} \to 0$ by assumption.

Let us first consider the case when $\mathcal{D} \to \infty$ as $n \to \infty$. Set $q_n := (d, d+1, d+1, \ldots, d+1, 1)$ as the architecture with constant width $d+1$ and depth $\mathcal{D} - 2$. Hanin and Sellke [2017] show that there exists a sequence of networks $\{g_n\}_{n \geq 1}$, each with architecture q_n, such that $\|g_n - f^*\|_\infty \to 0$.

One would like to have $g_n \in \mathcal{NN}_p^{\mathcal{B}}$, but g_n need not be bounded by $\mathcal{B}$. To solve this issue one can truncate g_n at level $\mathcal{B}$. Define the *truncation operator* $T_{\mathcal{B}}$ at level $\mathcal{B}$ by

$$T_{\mathcal{B}} x = \begin{cases} x & \text{if } |x| \leq \mathcal{B} \\ \mathcal{B} \cdot \operatorname{sign}(x) & \text{if } |x| > \mathcal{B} \end{cases} \quad \text{for all } x \in \mathbb{R}.$$

The truncation operator $T_{\mathcal{B}}$ is a piecewise linear function with two infliction points, it can therefore also be expressed as an MLP with one hidden layer. In fact,

$$T_{\mathcal{B}} x = \max(0, 2\mathcal{B} - \max(0, \mathcal{B} - x)) - \mathcal{B} = \sigma(2\mathcal{B} - \sigma(\mathcal{B} - x)) - \mathcal{B}.$$

The image of g_n is compact, and hence bounded, since the domain $[0,1]^d$ of g_n is compact. Therefore $l := \min_{x \in [0,1]^d} g_n(x)$ exists. Now define $g_n^* := T_{\mathcal{B}} \circ g_n$, denote the architecture of g_n^* by $q_n^* := (d, d+1, \ldots, d+1, 1, 1, 1)$, so $g_n^* \in \mathcal{NN}_{q_n^*}^{\mathcal{B}}$. Observe that g_n^* has depth $\mathcal{D}$ and $q_n^* \leq p$ since p has a width of at least $d+1$ at each layer.

By Lemma 2.7 we have $\|f^*\|_\infty = \|f_0 + \mu_\varepsilon\|_\infty \leq \mathcal{B}$. Since g_n^* is equal to g_n but truncated at level $\mathcal{B}$, we have

$$\|g_n^* - f^*\|_\infty \leq \|g_n - f^*\|_\infty \to 0.$$

Combining this with Proposition 2.5, we obtain

$$\begin{aligned} \lim_{n\to\infty} \inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|f - f^*\|_{L^2(\nu)}^2 &\leq \lim_{n\to\infty} \inf_{f \in \mathcal{NN}_{q_n^*}^{\mathcal{B}}} \|f - f^*\|_{L^2(\nu)}^2 \leq \lim_{n\to\infty} \|g_n^* - f^*\|_{L^2(\nu)}^2 \\ &= \lim_{n\to\infty} \int_{[0,1]^d} (g_n^* - f^*)^2 d\nu \leq \lim_{n\to\infty} \|g_n^* - f^*\|_\infty^2 \int_{[0,1]^d} d\nu \\ &= \lim_{n\to\infty} \|g_n^* - f^*\|_\infty^2 = 0. \end{aligned}$$

Now we consider the case when $\mathcal{W} \to \infty$ as $n \to \infty$. Hornik [1991] shows that there exists a sequence of MLP's $\{h_n\}_{n\geq 0}$ with one hidden layer and width $w_n \leq \mathcal{W}$ such that $\|h_n - f^*\|_\infty \to 0$ as $n \to \infty$. Denote the architecture of h_n by $a_n := (d, w_n, 1)$. Just like before we can truncate h_n to get a network $h_n^* := T_{\mathcal{B}} \circ h_n$ with architecture $(d, w_n, 1, 1, 1)$.

We know that $\mathcal{W} \to \infty$ as $n \to \infty$, where $\mathcal{W}$ is maximum width of p. Now for each n, denote

$$i_n = \arg \max_{1\leq i\leq \mathcal{D}} p_i,$$

where $p = (d, p_1, \ldots, p_{\mathcal{D}}, 1)$. Now we add identity layers in front of the hidden layer of h_n^* such that h_n^* has width w_n at hidden layer i_n. We can do this by setting the weight matrices to I_d and the bias vectors to 0. Hence we have written h_n^* using a network with architecture $a_n^* := (d, d, \ldots, d, w_n, 1, 1, 1)$. Observe that $a_n^* \leq p$, where we used the assumption that $\mathcal{D} \geq 3$ and $p_i \geq d + 1$ for all i. By Proposition 2.5, we obtain in the same way as before,

$$\lim_{n\to\infty} \inf_{f\in\mathcal{NN}_p^{\mathcal{B}}} \|f - f^*\|^2_{L^2(\nu)} \leq \lim_{n\to\infty} \inf_{f\in\mathcal{NN}^{\mathcal{B}}_{a_n^*}} \|f - f^*\|^2_{L^2(\nu)} \leq \lim_{n\to\infty} \|h_n^* - f^*\|^2_{L^2(\nu)} = 0.$$

Notice that if $\mathcal{W} \to \infty$ and $\mathcal{D} \to \infty$ as $n \to \infty$, the above proof for $\mathcal{W} \to \infty$ still works. Hence we have proven that

$$\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}} \{\mathcal{R}(f) - \mathcal{R}(f^*)\} = \inf_{f\in\mathcal{NN}_p^{\mathcal{B}}} \|f - f^*\|^2_{L^2(\nu)} \to 0 \quad \text{as} \quad n \to \infty,$$

under the assumption that $\mathcal{S} \to \infty$ as $n \to \infty$, which proves the result.

If $f_0 \in \mathcal{NN}_p^{\mathcal{B}}$, it follows that $f^* \in \mathcal{NN}_p^{\mathcal{B}}$. Therefore

$$\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right] \leq C_0\mathcal{B}^5(\log n)^5\frac{1}{n}\mathcal{SD}\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}} \|f - f^*\|^2_{L^2(\nu)} = C_0\mathcal{B}^5(\log n)^5\frac{1}{n}\mathcal{SD}\log\mathcal{S} \to 0$$

by assumption. This proves the final remark that if $f_0 \in \mathcal{NN}_p^{\mathcal{B}}$, the assumption that $\mathcal{S} \to \infty$ can be dropped. □

Under the assumption that $\mu_\varepsilon = 0$ and other slightly different assumptions, Theorem 3.8 has originally been proven in Jiao et al. [2023, Theorem 4.1]. They claim that the statement follows immediately from Yarotsky [2018, Theorem 1], which seems difficult to verify. They also state their version of Theorem 3.8 without assuming that each layer has width at least $d + 1$ and p has depth $D \geq 3$. In an unpublished but often cited article on arXiv, Hanin and Sellke [2017] prove in Theorem 1 that ReLU networks of constant width w are dense in $C([0, 1]^d, \mathbb{R})$ only for $w \geq d + 1$. This shows that the assumption on the minimum with of $d + 1$ is a necessary assumption which has been left out by Jiao et al. [2023, Theorem 4.1].

Our assumption that $\mathcal{D} \geq 3$, is purely one of convenience. It is used in the proof only to ensure the approximating network is bounded by $\mathcal{B}$. Without this assumption it is still true that the network is bounded by $\mathcal{B} + \eta$ for some small $\eta > 0$, since the network approximates f^* with maximum error η and f^* is bounded by $\mathcal{B}$.

Remember from Lemma 2.7 that for the square loss we have $f^* = f_0 + \mu_\varepsilon$. Theorem 3.8 shows that the ERM $\hat{f}_n$ converges to $f^* = f_0 + \mu_\varepsilon$ in expectation in $L^2(\nu)$ if ε is sub-exponential. If one assumes the contamination model introduced in Section 2.6, it clearly holds that $\mu_\varepsilon = \alpha\mu_{\text{outl}}$, where α is the mixture rate and μ_{outl} is the expectation of the outlier density.

In Section 3.1 it is shown that a normal distribution is sub-exponential. Furthermore, if the outlier density is sub-exponential, ε is sub-exponential as a mixture of sub-exponential random variables [see Proposition 3.3]. Hence by Theorem 3.8,

$$\hat{f}_n \to f_0 + \alpha\mu_{\text{outl}} \text{ in } L^2(\nu) \text{ in expectation.}$$

First of all, if there are no outliers, that is $\alpha = 0$, the ERM $\hat{f}_n$ is a consistent estimator for f_0. Whenever there are outliers, but the outlier density is sub-exponential and has zero mean, $\hat{f}_n$ is still a consistent estimator for f_0 despite the presence of outliers. The ERM is robust against mean zero, sub-exponential outliers. An example of a mean zero, sub-exponential outlier density is a normal distribution with higher variance and zero mean.

At the same time, if one has sub-exponential outliers that do not have zero mean, the ERM $\hat{f}_n$ will be biased. The larger the proportion of outliers, that is the mixture rate α, the larger the bias will become. Examples of such distributions are exponential, gamma and χ^2_n distributions.

4 Square loss with general error density.

In Lemma 3.7 and Theorem 3.8 from the previous section we have seen error bounds and convergence results for the square loss under the assumption that the error ε is sub-exponential. In this section we prove that similar results still hold for any distribution for ε as long as $\mathbb{E}[|\varepsilon|^{2+\delta}] < \infty$ for some $\delta > 0$.

4.1 Generalized square loss prediction error bounds and convergence.

The following is a generalized version of Lemma 3.7.

Lemma 4.1. *Fix $\delta > 0$ and consider the regression model in (1.1). Assume that $\mathbb{E}[|\varepsilon_i|^{2+\delta}] < \infty$ and $\|f_0\|_\infty \le \mathcal{B} - |\mu_\varepsilon|$ for $\mathcal{B} \ge 1$ and $\mu_\varepsilon := \mathbb{E}[\varepsilon_1]$. Let $\hat{f}_n \in \arg\min_{f\in\mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ denote the empirical risk minimizer (ERM) and $L(a,y) = (a-y)^2$ the square loss. Then, for all $n \ge 1$,*

$$\mathbb{E}_S\left[\mathcal{R}(f^*) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] \le c_0 n^{-\frac{\delta}{4+\delta}}\mathcal{B}^4 \log\mathcal{N}_n\left(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right),$$

for some constant $c_0 > 0$ independent of $n, d, \mathcal{D}, \mathcal{W}, \mathcal{U}, \mathcal{S}$ and $\mathcal{B}$. Furthermore, for $n \ge \operatorname{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$, we have

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right] &= \mathbb{E}_S\|\hat{f}_n - f^*\|^2_{L^2(\nu)} \\ &\le C_0\left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}} \mathcal{S}\mathcal{D}\mathcal{B}^5\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\{\mathcal{R}(f) - \mathcal{R}(f^*)\} \\ &\le C_0\left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}} \mathcal{S}\mathcal{D}\mathcal{B}^5\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\|f - f^*\|^2_{L^2(\nu)},\end{aligned}$$

where $C_0 > 0$ is some constant independent of $n, d, \mathcal{D}, \mathcal{W}, \mathcal{U}, \mathcal{S}$ and $\mathcal{B}$.

Proof. The proof is quite long, therefore the proof is given in its own section [see Section 4.2]. □

One should first note that the remark given about Lemma 3.7 still for this lemma.

Assuming that $f_0 \in \mathcal{NN}_p^{\mathcal{B}}$, which implies $f^* \in \mathcal{NN}_p^{\mathcal{B}}$, Lemma 3.7 gives a convergence rate of $O\left(\frac{(\log n)^5}{n}\right)$ for sub-exponential ε. Since any sub-exponential random variable has all moments finite, one could also apply above lemma with ”$\delta = \infty$”. This would result in a convergence rate of $O\left(\frac{\log n}{n}\right)$, which is similar to the rate given by Lemma 3.7.

In order to properly define the risk $\mathcal{R}$, that is, the risk is finite, one needs to assume $\mathbb{E}[|\varepsilon|^2] < \infty$. So if ε has second finite moment, one can define the risk. If ε has slightly more moments finite, that is, $\mathbb{E}[|\varepsilon|^{2+\delta}] < \infty$ for some $\delta > 0$, we obtain convergence of the ERM $\hat{f}_n$ of order $O\left(\left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}}\right)$ when assuming $f_0 \in \mathcal{NN}_p^{\mathcal{B}}$. This shows that Lemma 4.1 gives convergence in the mildest possible assumption ε. Also observe that the larger δ, the faster the convergence will be, which is natural.

Using above lemma, Theorem 3.8 has a generalized formulation where ε is assumed to have $(2+\delta)$-th finite moment for some $\delta > 0$ instead of being sub-exponential. This generalization is given in the following theorem.

Theorem 4.2 (consistency of ERM under square loss)**.** *Fix $\delta > 0$ and consider the regression model from (1.1). Assume $\mathbb{E}[|\varepsilon_i|^{2+\delta}] < \infty$, f_0 is continuous on $[0,1]^d$ and $\|f_0\|_\infty \leq \mathcal{B} - |\mu_\varepsilon|$ with $\mathcal{B} \geq 1$ and $\mu_\varepsilon := \mathbb{E}[\varepsilon_1]$. Let $\hat{f}_n \in \arg\min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ denote the empirical risk minimizer over $\mathcal{NN}_p^{\mathcal{B}}$ and $L(a,y) = (a-y)^2$ the square loss. Also assume that each layer of p has at least width $d+1$ and p has depth $\mathcal{D} \geq 3$. If the architecture p satisfies,*

$$\mathcal{S} \to \infty \quad \text{and} \quad \left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}} \mathcal{S}\mathcal{D}\mathcal{B}^5 \log \mathcal{S} \to 0 \quad \text{as} \quad n \to \infty.$$

Then, the prediction error of the ERM $\hat{f}_n$ satisfies

$$\lim_{n\to\infty} \mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right] = \lim_{n\to\infty} \mathbb{E}_S\left[\|\hat{f}_n - f^*\|^2_{L^2(\nu)}\right] = 0.$$

Furthermore, if $f_0 \in \mathcal{NN}_p^{\mathcal{B}}$, the condition that $\mathcal{S} \to \infty$ can be dropped.

Proof. The proof is exactly the same is the proof of Theorem 3.8, while replacing Lemma 3.7 with Lemma 4.1. □

Just like we remarked after Theorem 3.8, if the contamination model from Section 2.6 is assumed for ε, we have, under the assumption that the outlier density has $(2+\delta)$-th finite moment,

$$\hat{f}_n \to f_0 + \alpha\mu_{\text{outl}} \text{ in } L^2(\nu) \text{ in expectation.}$$

Where α is the proportion of outliers, or the mixture rate, and μ_{outl} is the expectation of the outlier density.

Many outlier densities used to model outliers are not sub-exponential. Hence Theorem 3.8 is not applicable. However, our generalization assumes much less about the outlier density. Hence outlier densities like a t-distribution or a Fréchet distribution can also be used with some conditions on their parameters [see Table 3].

Theorem 4.2 shows that when there are no outliers, $\hat{f}_n$ is a consistent estimator for f_0. Conversely, when there are outliers with non-zero mean, $\hat{f}_n$ will be a biased estimator for f_0 with bias $\alpha\mu_{\text{outl}}$. If there are outliers, but with zero mean, $\hat{f}_n$ is still a consistent estimator for f_0. This shows that $\hat{f}_n$ is robust against zero mean outliers, such as outliers following a t-distribution.

4.2 Proof of Lemma 4.1.

In the following we present the proof of Lemma 4.1. The proof is an adaptation of the proof of Lemma 3.2 from Jiao et al. [2023]. In their work they assumed ε to be sub-exponential. We work under the milder assumption that $\mathbb{E}\left[|\varepsilon_i|^{2+\delta}\right] < \infty$. This change in assumption results in a different bound in the proof. However, many parts are still the same because they do not require any assumption on ε. For these parts, we have provided more details than Jiao et al. [2023] did to make the proof easier to follow. Let us first consider the case when $\mathbb{E}[\varepsilon] = 0$.

Let $S := \{Z_i\}_{i=1}^n := \{(X_i, Y_i)\}_{i=1}^n$ be a random sample of i.i.d. observations with distribution $Z = (X, Y)$ and $S' := \{Z_i'\}_{i=1}^n := \{(X_i', Y_i')\}_{i=1}^n$ be another sample of i.i.d. observations with distribution Z that is independent of S. Define

$$g(f, Z_i) := (f(X_i) - Y_i)^2 - (f_0(X_i) - Y_i)^2$$

for any f and observation Z_i. Now

$$\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] = \mathbb{E}_S\left[\mathbb{E}_Z(Y - f_0(X))^2 - 2\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{f}_n(X_i))^2 + \mathbb{E}_Z(Y - \hat{f}_n(X))^2\right]$$

$$= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{-2g(\hat{f}_n, Z_i) - 2(f_0(X_i) - Y_i)^2\right\} + \mathbb{E}_Z(Y - f_0(X))^2 + \mathbb{E}_Z(Y - \hat{f}_n(X))^2\right],$$

since Z_i' is independent from S and Z_i' has distribution Z we have

$$= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{-2g(\hat{f}_n, Z_i) - 2(f_0(X_i) - Y_i)^2 + \mathbb{E}_{Z_i'}\left((Y_i' - f_0(X_i'))^2 + (Y_i' - \hat{f}_n(X_i'))^2\right)\right\}\right],$$

taking expectation with respect to the whole sample S' instead of just Z_i' gives

$$= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{-2g(\hat{f}_n, Z_i) - 2(f_0(X_i) - Y_i)^2 + \mathbb{E}_{S'}\left(g(\hat{f}_n, Z_i') + 2(Y_i' - f_0(X_i'))^2\right)\right\}\right]$$

$$= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{-2g(\hat{f}_n, Z_i) + \mathbb{E}_{S'}g(\hat{f}_n, Z_i')\right\}\right] + \frac{1}{n}\sum_{i=1}^n \left(-2\mathbb{E}_S(f_0(X_i) - Y_i)^2 + 2\mathbb{E}_{S'}(f_0(X_i') - Y_i')^2\right)$$

$$= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{-2g(\hat{f}_n, Z_i) + \mathbb{E}_{S'}g(\hat{f}_n, Z_i')\right\}\right] \text{ since } S \text{ and } S' \text{ are independent.}$$

Thus we have derived that

$$\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{-2g(\hat{f}_n, Z_i) + \mathbb{E}_{S'}g(\hat{f}_n, Z_i')\right\}\right]. \tag{4.1}$$

Using Lemma 2.8 we can bound the prediction error with

$$\mathbb{E}_S\left[\|\hat{f}_n - f_0\|_{L^2(\nu)}^2\right] \le \mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] + 2\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}} \|\hat{f}_n - f_0\|_{L^2(\nu)}^2. \tag{4.2}$$

The remaining part of the proof focuses on giving an upper bound for $\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right]$.

For a shorter notation define

$$G(f, Z_i) := \mathbb{E}_{S'}\left[g(f, Z_i')\right] - 2g(f, Z_i) \quad \text{for all } i \text{ and } f \in \mathcal{NN}_p^{\mathcal{B}}.$$

Observe that $\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G(\hat{f}_n, Z_i)\right]$ by above definition and (4.1).

Let $\beta_n \ge \mathcal{B} \ge 1$ be a positive constant depending on the sample size n. Define the *truncation operator* T_{β_n} at level β_n by

$$T_{\beta_n}x = \begin{cases} x & \text{if } |x| \le \beta_n \\ \beta_n \cdot \operatorname{sign}(x) & \text{if } |x| > \beta_n \end{cases} \quad \text{for all } x \in \mathbb{R}.$$

Set $f_{\beta_n}(x) := \mathbb{E}\left[T_{\beta_n}Y \,\middle|\, X = x\right]$ for all x as the regression function of the truncated Y. In similar fashion as before we define

$$g_{\beta_n}(f, Z_i) := (f(X_i) - T_{\beta_n}Y_i)^2 - (f_{\beta_n}(X_i) - T_{\beta_n}Y_i)^2 \text{ and}$$

$$G_{\beta_n}(f,Z_i) := \mathbb{E}_{S'}\left[g_{\beta_n}(f,Z_i')\right] - 2g_{\beta_n}(f,Z_i) \ \text{ for all } i \text{ and } f \in \mathcal{NN}_p^{\mathcal{B}}.$$

Then for any $f \in \mathcal{NN}_p^{\mathcal{B}}$ and any i we have

$$\begin{aligned}
&\left|g(f,Z_i) - g_{\beta_n}(f,Z_i)\right| = \left|(f(X_i)-Y_i)^2 - (f_0(X_i)-Y_i)^2 - (f(X_i)-T_{\beta_n}Y_i)^2 + (f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2\right| \\
&= \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - 2f(X_i)Y_i - f_0(X_i)^2 + 2f_0(X_i)Y_i + 2f(X_i)T_{\beta_n}Y_i - (T_{\beta_n}Y_i)^2\right| \\
&= \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - f_0(X_i)^2 - 2Y_i(f(X_i)-f_0(X_i)) + 2f(X_i)T_{\beta_n}Y_i - (T_{\beta_n}Y_i)^2\right| \\
&= \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - (f_0(X_i)-T_{\beta_n}Y_i)^2 - 2f_0(X_i)T_{\beta_n}Y_i - 2Y_i(f(X_i)-f_0(X_i)) + 2f(X_i)T_{\beta_n}Y_i\right| \\
&= \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - (f_0(X_i)-T_{\beta_n}Y_i)^2 + (f(X_i)-f_0(X_i))\cdot 2T_{\beta_n}Y_i - 2Y_i(f(X_i)-f_0(X_i))\right| \\
&= \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - (f_0(X_i)-T_{\beta_n}Y_i)^2 + 2(T_{\beta_n}Y_i - Y_i)(f(X_i)-f_0(X_i))\right| \\
&\le \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - (f_0(X_i)-T_{\beta_n}Y_i)^2\right| + 2\left|f(X_i)-f_0(X_i)\right|\cdot\left|T_{\beta_n}Y_i - Y_i\right|.
\end{aligned}$$

Since $\|f_0\|_\infty \le \mathcal{B}$ and $\|f\|_\infty \le \mathcal{B}$ we get $\left|f(X_i)-f_0(X_i)\right| \le \|f_0\|_\infty + \|f\|_\infty \le 2\mathcal{B}$. Thus

$$\begin{aligned}
&\le \left|(f_{\beta_n}(X_i)-T_{\beta_n}Y_i)^2 - (f_0(X_i)-T_{\beta_n}Y_i)^2\right| + 4\mathcal{B}\left|T_{\beta_n}Y_i - Y_i\right| \\
&= \left|f_{\beta_n}(X_i)^2 - 2f_{\beta_n}(X_i)T_{\beta_n}Y_i - f_0(X_i)^2 + 2f_0(X_i)T_{\beta_n}Y_i\right| + 4\mathcal{B}\left|T_{\beta_n}Y_i - Y_i\right| \\
&= \left|(f_{\beta_n}(X_i)+f_0(X_i))(f_{\beta_n}(X_i)-f_0(X_i)) - 2T_{\beta_n}Y_i(f_{\beta_n}(X_i)-f_0(X_i))\right| + 4\mathcal{B}\left|T_{\beta_n}Y_i - Y_i\right| \\
&\le \left|f_{\beta_n}(X_i)-f_0(X_i)\right|\cdot\left|f_{\beta_n}(X_i)+f_0(X_i)-2T_{\beta_n}Y_i\right| + 4\mathcal{B}\left|T_{\beta_n}Y_i - Y_i\right|.
\end{aligned}$$

Note that $f_{\beta_n}(X_i) \le \beta_n$, $f_0(X_i) \le \mathcal{B} \le \beta_n$ and $T_{\beta_n}Y_i \le \beta_n$. Therefore

$$\left|f_{\beta_n}(X_i)+f_0(X_i)-2T_{\beta_n}Y_i\right| \le 4\beta_n,$$

which implies

$$\left|g(f,Z_i) - g_{\beta_n}(f,Z_i)\right| \le 4\beta_n\left|f_{\beta_n}(X_i)-f_0(X_i)\right| + 4\mathcal{B}\left|T_{\beta_n}Y_i - Y_i\right|.$$

Notice that

$$\left|T_{\beta_n}Y_i - Y_i\right| = \begin{cases} 0 & \text{if } |Y_i| \le \beta_n \\ \left|\beta_n\cdot\operatorname{sign}(Y_i) - Y_i\right| & \text{if } |Y_i| > \beta_n \end{cases}.$$

if $Y_i > \beta_n$ then $\left|\beta_n\cdot\operatorname{sign}(Y_i) - Y_i\right| = Y_i - \beta_n \le |Y_i|$. Similarly $\left|\beta_n\cdot\operatorname{sign}(Y_i) - Y_i\right| = \beta_n - Y_i = |Y_i| - \beta_n \le |Y_i|$ whenever $Y_i < -\beta_n$. Thus it holds in general that

$$\left|T_{\beta_n}Y_i - Y_i\right| \le \begin{cases} 0 & \text{if } |Y_i| \le \beta_n \\ |Y_i| & \text{if } |Y_i| > \beta_n \end{cases},$$

or written more simply as $\left|T_{\beta_n}Y_i - Y_i\right| \le |Y_i|\mathbb{1}_{\{|Y_i|>\beta_n\}}$. It follows that

$$\left|g(f,Z_i) - g_{\beta_n}(f,Z_i)\right| \le 4\beta_n\left|f_{\beta_n}(X_i)-f_0(X_i)\right| + 4\mathcal{B}|Y_i|\mathbb{1}_{\{|Y_i|>\beta_n\}}.$$

By conditional Jensen's inequality:

$$\begin{aligned}
\left|f_{\beta_n}(X_i)-f_0(X_i)\right| &= \left|\mathbb{E}\left[T_{\beta_n}Y_i \,\middle|\, X = X_i\right] - \mathbb{E}\left[Y_i \,\middle|\, X = X_i\right]\right| \\
&= \left|\mathbb{E}\left[T_{\beta_n}Y_i - Y_i \,\middle|\, X = X_i\right]\right| \\
&\le \mathbb{E}\left[\left|T_{\beta_n}Y_i - Y_i\right| \,\middle|\, X = X_i\right],
\end{aligned}$$

thus $\mathbb{E}_S\left(\left|f_{\beta_n}(X_i)-f_0(X_i)\right|\right) \le \mathbb{E}_S\left(\mathbb{E}\left[\left|T_{\beta_n}Y_i - Y_i\right| \,\middle|\, X = X_i\right]\right) = \mathbb{E}_S\left(\left|T_{\beta_n}Y_i - Y_i\right|\right)$. Combining this with the above result yields

$$\begin{aligned}
\mathbb{E}_S\left[g(f,Z_i)\right] &= \mathbb{E}_S\left[g(f,Z_i) - g_{\beta_n}(f,Z_i) + g_{\beta_n}(f,Z_i)\right] \\
&\leq \mathbb{E}_S\left[\left|g(f,Z_i) - g_{\beta_n}(f,Z_i)\right| + g_{\beta_n}(f,Z_i)\right] \\
&\leq \mathbb{E}_S\left[4\beta_n\left|f_{\beta_n}(X_i) - f_0(X_i)\right| + 4\mathcal{B}|Y_i|\mathbb{1}_{\{|Y_i|>\beta_n\}} + g_{\beta_n}(f,Z_i)\right] \\
&\leq \mathbb{E}_S\left[g_{\beta_n}(f,Z_i)\right] + 4\beta_n\mathbb{E}_S\left(\left|f_{\beta_n}(X_i) - f_0(X_i)\right|\right) + 4\beta_n\mathbb{E}_S\left[|Y_i|\mathbb{1}_{\{|Y_i|>\beta_n\}}\right] \\
&\leq \mathbb{E}_S\left[g_{\beta_n}(f,Z_i)\right] + 8\beta_n\mathbb{E}_S\left[|Y_i|\mathbb{1}_{\{|Y_i|>\beta_n\}}\right] \\
&\leq \mathbb{E}_S\left[g_{\beta_n}(f,Z_i)\right] + 8\beta_n\mathbb{E}_S\left[|Y_i|\frac{|Y_i|^{1+\delta}}{\beta_n^{1+\delta}}\right] \\
&= \mathbb{E}_S\left[g_{\beta_n}(f,Z_i)\right] + 8\frac{\mathbb{E}\left[|Y_i|^{2+\delta}\right]}{\beta_n^{\delta}} \qquad (4.3)
\end{aligned}$$

Note that f_0 being continuous implies that $f_0(X_i)$ is bounded and therefore sub-exponential by Corollary 3.6.1 since X_i is bounded. Now for all i

$$\mathbb{E}\left[|Y_i|^{2+\delta}\right] = \mathbb{E}\left[|f_0(X_i)+\varepsilon_i|^{2+\delta}\right] \leq 2^{1+\delta}\left(\mathbb{E}\left[|f_0(X_i)|^{2+\delta}\right] + \mathbb{E}\left[|\varepsilon_i|^{2+\delta}\right]\right) < \infty,$$

by the C_r-inequality and using the fact that $f_0(X_i)$ is sub-exponential and $\mathbb{E}[|\varepsilon_i|^{2+\delta}] < \infty$ by assumption. Analogously it holds that

$$\begin{aligned}
\mathbb{E}_S\left[g_{\beta_n}(f,Z_i)\right] &\leq \mathbb{E}_S\left[\left|g(f,Z_i) - g_{\beta_n}(f,Z_i)\right| + g(f,Z_i)\right] \\
&\leq \mathbb{E}_S\left[g(f,Z_i)\right] + 8\beta_n\mathbb{E}_S\left[|Y_i|\mathbb{1}_{\{|Y_i|>\beta_n\}}\right] \\
&\leq \mathbb{E}_S\left[g(f,Z_i)\right] + 8\frac{\mathbb{E}\left[|Y_i|^{2+\delta}\right]}{\beta_n^{\delta}}.
\end{aligned}$$

Using this in combination with (4.3) we obtain that

$$\begin{aligned}
\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] &= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G(\hat{f}_n, Z_i)\right] \\
&= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}_{S'}g(\hat{f}_n, Z_i') - 2g(\hat{f}_n, Z_i)\right\}\right] \\
&= \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}_S\mathbb{E}_{S'}\left[g(\hat{f}_n, Z_i')\right] - 2\mathbb{E}_S\left[g(\hat{f}_n, Z_i)\right]\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}_S\mathbb{E}_{S'}\left[g_{\beta_n}(\hat{f}_n, Z_i')\right] + 8\frac{\mathbb{E}\left[|Y_i|^{2+\delta}\right]}{\beta_n^{\delta}} - 2\mathbb{E}_S\left[g(\hat{f}_n, Z_i)\right]\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_S\mathbb{E}_{S'}\left[g_{\beta_n}(\hat{f}_n, Z_i')\right] + 8\frac{\mathbb{E}\left[|Y_i|^{2+\delta}\right]}{\beta_n^{\delta}} - 2\mathbb{E}_S\left[g_{\beta_n}(\hat{f}_n, Z_i)\right] \\
&\quad + 16\frac{\mathbb{E}\left[|Y_i|^{2+\delta}\right]}{\beta_n^{\delta}} \\
&= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\right] + 24\beta_n^{-\delta}\mathbb{E}\left[|Y_i|^{2+\delta}\right] \\
&= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\right] + c_1\beta_n^{-\delta}, \qquad (4.4)
\end{aligned}$$

where $c_1 := 24\mathbb{E}\left[|Y_i|^{2+\delta}\right]$ is a constant not depending on β_n and n.

Observing that for any $f \in \mathcal{NN}_p^{\mathcal{B}}$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f, Z_i) &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{S'}\left[g_{\beta_n}(f, Z_i')\right] - 2g_{\beta_n}(f, Z_i)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S'}\left[g_{\beta_n}(f, Z_i')\right] - \frac{2}{n}\sum_{i=1}^{n} g_{\beta_n}(f, Z_i)\\ &= \mathbb{E}_{S'}[g_{\beta_n}(f, Z_1')] - \frac{2}{n}\sum_{i=1}^{n} g_{\beta_n}(f, Z_i),\end{aligned}$$

it follows that the tail of $\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(\hat{f}_n, Z_i)$ is bounded as

$$\begin{aligned}&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(\hat{f}_n, Z_i) > t\right\}\\ &\leq \mathbb{P}\left\{\exists f \in \mathcal{NN}_p^{\mathcal{B}} : \frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(\hat{f}_n, Z_i) > t\right\}\\ &= \mathbb{P}\left\{\exists f \in \mathcal{NN}_p^{\mathcal{B}} : \mathbb{E}_{Z_1'}[g_{\beta_n}(f, Z_1')] - \frac{2}{n}\sum_{i=1}^{n} g_{\beta_n}(f, Z_i) > t\right\}\\ &= \mathbb{P}\Bigg\{\exists f \in \mathcal{NN}_p^{\mathcal{B}} : \mathbb{E}\left[(f(X) - T_{\beta_n}Y)^2 - (f_{\beta_n}(X) - T_{\beta_n}Y)^2\right] - \frac{2}{n}\sum_{i=1}^{n}\Big\{(f(X_i) - T_{\beta_n}Y_i)^2 - (f_{\beta_n}(X_i)\\ &\quad - T_{\beta_n}Y_i)^2\Big\} > t\Bigg\}\\ &= \mathbb{P}\Bigg\{\exists f \in \mathcal{NN}_p^{\mathcal{B}} : \frac{1}{2}\mathbb{E}|f(X) - T_{\beta_n}Y|^2 - \frac{1}{2}\mathbb{E}|f_{\beta_n}(X) - T_{\beta_n}Y|^2 - \frac{1}{n}\sum_{i=1}^{n}\Big\{(f(X_i) - T_{\beta_n}Y_i)^2 - (f_{\beta_n}(X_i)\\ &\quad - T_{\beta_n}Y_i)^2\Big\} > \frac{1}{2}t\Bigg\}\\ &= \mathbb{P}\Bigg\{\exists f \in \mathcal{NN}_p^{\mathcal{B}} : \mathbb{E}|f(X) - T_{\beta_n}Y|^2 - \mathbb{E}|f_{\beta_n}(X) - T_{\beta_n}Y|^2 - \frac{1}{n}\sum_{i=1}^{n}\Big\{(f(X_i) - T_{\beta_n}Y_i)^2 - (f_{\beta_n}(X_i)\\ &\quad - T_{\beta_n}Y_i)^2\Big\} > \frac{1}{2}\left(t + \mathbb{E}|f(X) - T_{\beta_n}Y|^2 - \mathbb{E}|f_{\beta_n}(X) - T_{\beta_n}Y|^2\right)\Bigg\} = (*).\end{aligned}$$

Since $|T_{\beta_n}Y| \leq \beta_n$, $\beta_n \geq 1$ and $\|f\|_\infty \leq \mathcal{B} \leq \beta_n$ for all $f \in \mathcal{NN}_p^{\mathcal{B}}$ we can apply Theorem 11.4 from Györfi et al. [2002, see p. 201] with $\varepsilon = \frac{1}{2}$ and $\alpha = \beta = \frac{t}{2}$ so that

$$(*) \leq 14\mathcal{N}_n\left(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right)\exp\left(-\frac{\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{2}tn}{214\cdot\frac{3}{2}\beta_n^4}\right) = 14\mathcal{N}_n\left(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right)\exp\left(-\frac{tn}{5136\beta_n^4}\right).$$

Using the fact that for any real-valued random variable X we can write

$$\mathbb{E}[X] = \int_0^\infty 1 - F_X(x)dx - \int_{-\infty}^0 F_X(x)dx \leq \int_0^\infty 1 - F_X(x)dx,$$

we obtain for any $a_n > 0$ that:

$$\begin{aligned}\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\right] &\leq \int_0^\infty \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i) > t\right\} dt \\ &\leq \int_0^{a_n} 1dt + \int_{a_n}^\infty \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i) > t\right\} dt \\ &\leq a_n + \int_{a_n}^\infty 14\mathcal{N}_n\left(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right)\exp\left(-\frac{tn}{5136\beta_n^4}\right) dt.\end{aligned}$$

Observe that for $a, b \in \mathbb{R}_{>0}$ with $a \geq b$ it holds that $\mathcal{N}_n(a, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) \leq \mathcal{N}_n(b, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})$ since there are less balls needed to cover the space when the radius of the balls is increased. Therefore

$$\begin{aligned}\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\right] &\leq a_n + \int_{a_n}^\infty 14\mathcal{N}_n\left(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right)\exp\left(-\frac{tn}{5136\beta_n^4}\right) dt \\ &\leq a_n + \int_{a_n}^\infty 14\mathcal{N}_n\left(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right)\exp\left(-\frac{tn}{5136\beta_n^4}\right) dt \\ &= a_n + 14\mathcal{N}_n\left(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right)\exp\left(-\frac{a_n n}{5136\beta_n^4}\right)\frac{5136\beta_n^4}{n}.\end{aligned}$$

Choose $a_n = \frac{5136\beta_n^4}{n}\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}))$. Note that

$$\begin{aligned}\frac{a_n}{80\beta_n} &= \frac{5136\beta_n^4}{n}\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}))\frac{1}{80\beta_n} = \log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}))\frac{321}{5}\beta_n^3\frac{1}{n} \\ &\geq \log(14)\frac{321}{5}\frac{1}{n} \geq \frac{1}{n},\end{aligned}$$

and therefore $\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) \geq \mathcal{N}_n(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})$. From this we derive

$$\begin{aligned}\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\right] &\leq \log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}))\frac{5136\beta_n^4}{n} \\ &\quad + 14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})\exp\left(-\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}))\right)\frac{5136\beta_n^4}{n} \\ &= \frac{5136\beta_n^4}{n}\left(\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right).\end{aligned}$$

Combining this with (4.4) we obtain that

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] &\leq \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\right] + c_1\beta_n^{-\delta} \\ &\leq \frac{5136\beta_n^4}{n}\left(\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + c_1\beta_n^{-\delta}.\end{aligned} \tag{4.5}$$

Choose $\beta_n = \mathcal{B} n^{\frac{1}{4+\delta}} \geq \mathcal{B}$, then

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] &\leq 5136\mathcal{B}^4 n^{-\frac{\delta}{4+\delta}}\left(\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + c_1\mathcal{B}^{-\delta} n^{-\frac{\delta}{4+\delta}} \\ &= n^{-\frac{\delta}{4+\delta}}\left[5136\mathcal{B}^4\left(\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + c_1\mathcal{B}^{-\delta}\right] \\ &\leq n^{-\frac{\delta}{4+\delta}}\left[5136\mathcal{B}^4\left(\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + c_1\right]. \\ &\leq n^{-\frac{\delta}{4+\delta}}\left[2\cdot 5136\mathcal{B}^4 \log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + c_1\right]\end{aligned}$$

Now observe that for any positive, strictly increasing function $h : \mathbb{N} \to \mathbb{R}$, we have

$$k_1 h(n) + k_2 = k_1 h(n) + h(1)\frac{k_2}{h(1)} \leq h(n)\left(k_1 + \frac{k_2}{h(1)}\right) = Kh(n),$$

for any constants k_1 and k_2 and $K := k_1 + \frac{k_2}{h(1)}$. Applying this to above inequality results in

$$\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] \leq c_0 n^{-\frac{\delta}{4+\delta}}\mathcal{B}^4 \log\mathcal{N}_n\left(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}\right),$$

for some constant $c_0 > 0$. This proves the first part of the lemma for $\mathbb{E}[\varepsilon] = 0$.

Lastly, the uniform covering number can be bounded by the pseudo dimension $\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$ of $\mathcal{NN}_p^{\mathcal{B}}$. Which in turn can be bounded by properties of the function class $\mathcal{NN}_p^{\mathcal{B}}$. By Theorem 12.2 in Anthony and Bartlett [1999], for any $n \geq \mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$,

$$\mathcal{N}_n(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) \leq \left(\frac{en\mathcal{B}}{n^{-1}\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})}\right)^{\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})} = \left(\frac{en^2\mathcal{B}}{\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})}\right)^{\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})}.$$

Moreover, by Bartlett et al. [2019], there exists a constant $C > 0$ such that

$$\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}}) \leq C \cdot \mathcal{SD}\log(\mathcal{S}).$$

Using the fact that $\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}}) \geq 1$, we obtain

$$\begin{aligned}\log\mathcal{N}_n(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) &\leq \mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})\left(\log(e\mathcal{B}n^2) - \log\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})\right) \\ &\leq C \cdot \mathcal{SD}\log(\mathcal{S})(\log(e\mathcal{B}n^2)) \\ &= C \cdot \mathcal{SD}\log(\mathcal{S})\left(2\log(n) + \log(e) + \log(\mathcal{B})\right).\end{aligned}$$

It is easy to check that $2\log(n) + \log(\mathcal{B}) \leq 2\mathcal{B}\log(n)$ for all $n \geq 1$. Furthermore, clearly $\log(e) \leq 2\mathcal{B}\log(n)$. Therefore

$$\log\mathcal{N}_n(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) \leq 4C \cdot \mathcal{SD}\log(\mathcal{S})\mathcal{B}\log(n).$$

Combining this with (4.5) for general $\beta_n \geq \mathcal{B}$, we obtain

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] &\leq \frac{5136\beta_n^4}{n}\left(\log(14\mathcal{N}_n(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + c_1\beta_n^{-\delta} \\ &\leq \frac{5136\beta_n^4}{n}\left(4C \cdot \mathcal{SD}\log(\mathcal{S})\mathcal{B}\log(n) + \log(14) + 1\right) + c_1\beta_n^{-\delta} \\ &\leq \frac{5136\beta_n^4}{n}\left(12C \cdot \mathcal{SD}\log(\mathcal{S})\mathcal{B}\log(n)\right) + c_1\beta_n^{-\delta}.\end{aligned}$$

Choose $\beta_n = \mathcal{B}\left(\frac{n}{\log n}\right)^{\frac{1}{4+\delta}} \geq \mathcal{B}$, then

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(f_0) - 2\mathcal{R}_n(\hat{f}_n) + \mathcal{R}(\hat{f}_n)\right] &\leq 12C \cdot 5136\mathcal{S}\mathcal{D}\log(\mathcal{S})\mathcal{B}^5\left(\frac{n}{\log n}\right)^{\frac{4}{4+\delta}-1} + c_1\mathcal{B}^{-\delta}\left(\frac{n}{\log n}\right)^{\frac{-\delta}{4+\delta}} \\ &= \left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}}\left[12C \cdot 5136\mathcal{S}\mathcal{D}\log(\mathcal{S})\mathcal{B}^5 + c_1\left(\frac{1}{\mathcal{B}}\right)^{\delta}\right] \\ &\leq \left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}}\left[12C \cdot 5136\mathcal{S}\mathcal{D}\log(\mathcal{S})\mathcal{B}^5 + c_1\right] \\ &\leq \left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}}\left[12C \cdot 5136\mathcal{S}\mathcal{D}\log(\mathcal{S})\mathcal{B}^5 + c_1\mathcal{S}\mathcal{D}\log(\mathcal{S})\mathcal{B}^5\right] \\ &\leq C_0\left(\frac{\log n}{n}\right)^{\frac{\delta}{4+\delta}}\mathcal{S}\mathcal{D}\mathcal{B}^5\log\mathcal{S}, \qquad (4.6)\end{aligned}$$

for some constant $C_0 > 0$. This gives the bound on the prediction error for $\mathbb{E}[\varepsilon] = 0$ when combined with (4.2) and Lemma 2.8.

The result can be generalized to $\mathbb{E}[\varepsilon] \neq 0$ in the same way as has been done in the proof of Theorem 3.8. □

5 Error bound with Lipschitz loss.

In this section we consider a general loss function that is Lipschitz continuous [see Section 2.5]. Similar results to the ones from Section 4 are formulated, with the goal in mind to allow for more robust estimation of the regression function.

5.1 Lipschitz loss prediction error bounds and convergence.

Let us first formulate a result similar to Lemma 3.7 and Lemma 4.1.

Lemma 5.1. *Consider the regression model in (1.1) with unknown regression function f_0 and target function f^* defined in (2.3). Assume that the loss function is λ_L-Lipschitz in both its arguments, $L(x,x) = 0$ for all $x \in \mathbb{R}$, $\mathbb{E}[|\varepsilon|^p] < \infty$ for some $p > 1$ and $\|f^*\|_\infty \leq \mathcal{B}$ with $\mathcal{B} \geq 1$. Let $\hat{f}_n \in \arg\min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ denote the empirical risk minimizer (ERM). Then, for $n \geq \frac{1}{2}\mathrm{Pdim}(\mathcal{NN}_p^{\mathcal{B}})$,*

$$\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right] \leq c_0\left(\frac{\lambda_L\mathcal{B}}{n^{1-\frac{1}{p}}}\right)\log\mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) + 2\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}}\{\mathcal{R}(f) - \mathcal{R}(f^*)\},$$

where $c_0 > 0$ is a constant independent of $n, d, \lambda_L, \mathcal{B}, \mathcal{S}, \mathcal{W}$ and $\mathcal{D}$. Furthermore,

$$\mathbb{E}_S\left\{\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right\} \leq C_0\left(\frac{\log n}{n^{1-\frac{1}{p}}}\right)\lambda_L\mathcal{B}\mathcal{D}\mathcal{S}\log\mathcal{S} + 2\inf_{f \in \mathcal{NN}_p^{\mathcal{B}}}\left\{\mathcal{R}(f) - \mathcal{R}(f^*)\right\},$$

where $C_0 > 0$ is a constant independent of $n, d, \lambda_L, \mathcal{B}, \mathcal{S}, \mathcal{W}$ and $\mathcal{D}$.

It is important to note that Lemma 5.1 is not a generalization of Lemma 4.1 since the square loss is not Lipschitz continuous. Hence one cannot be stated to be better than the other by comparing convergence rates, since the loss functions considered are different.

Under the assumption that $f^* \in \mathcal{NN}_p^{\mathcal{B}}$, a convergence rate of $O\left(\frac{\log n}{n^{1-\frac{1}{p}}}\right)$ is obtained by above lemma. When this is not the case, the right term in the upper bound will not vanish for a fixed architecture and one will remain with an approximation error. Observe that when ε has more moments finite, the convergence rate becomes faster, which is natural. When all moments are finite we obtain a rate of $O\left(\frac{\log n}{n}\right)$, which agrees with the rate for the square loss in Lemma 4.1 with all moments finite.

Finally, we show a similar consistency result to Theorem 4.2.

Theorem 5.2 (consistency of ERM under Lipschitz loss)**.** *Consider the regression model in (1.1) with unknown regression function f_0. Assume that the loss function L is λ_L-Lipschitz in both its arguments, $L(x,x) = 0$ for all $x \in \mathbb{R}$, $\mathbb{E}[|\varepsilon|^p] < \infty$ for some $p > 1$ and, f^* is continuous on $[0,1]^d$ and $\|f^*\|_\infty \leq \mathcal{B}$ with $\mathcal{B} \geq 1$. Let $\hat{f}_n \in \arg\min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}_n(f)$ denote the empirical risk minimizer over $\mathcal{NN}_p^{\mathcal{B}}$. Also assume that each layer of p has at least width $d+1$ and p has depth $\mathcal{D} \geq 3$. If the architecture p satisfies,*

$$\mathcal{S} \to \infty \quad \text{and} \quad \left(\frac{\log n}{n^{1-\frac{1}{p}}}\right) \mathcal{B}\mathcal{D}\mathcal{S}\log\mathcal{S} \to 0 \quad \text{as} \quad n \to \infty.$$

Then, the prediction error of the ERM $\hat{f}_n$ satisfies

$$\lim_{n\to\infty} \mathbb{E}_S\left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right] = 0.$$

Furthermore, if $f^* \in \mathcal{NN}_p^{\mathcal{B}}$, the condition that $\mathcal{S} \to \infty$ can be dropped.

Proof. The proof is similar to the proof of Theorem 3.8. The big difference is that we cannot write $\mathcal{R}(f) - \mathcal{R}(f^*)$ as a norm, which was possible for the square loss. Luckily, one can upper bound the difference by the L^1 norm. For any f,

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f^*) &= \mathbb{E}_Z\left[L(f(X),Y) - L(f^*(X),Y)\right] \leq \mathbb{E}_Z\left[\big|L(f(X),Y) - L(f^*(X),Y)\big|\right] \\ &\leq \mathbb{E}_X\left[\lambda_L\big|f(X) - f^*(X)\big|\right] =: \lambda_L\|f - f^*\|_{L^1(\nu)}.\end{aligned}$$

Hence it follows that

$$\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\{\mathcal{R}(f) - \mathcal{R}(f^*)\} \leq \lambda_L \inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\|f - f^*\|_{L^1(\nu)}.$$

Above inequality was originally stated by Shen et al. [2021, Lemma 3.2].

The rest of the proof is exactly the same as the proof of Theorem 3.8, while using the fact that convergence in L^∞ implies convergence in L^1 in combination with Lemma 5.1. □

By definition of f^* any estimator f will have a risk greater or equal to $\mathcal{R}(f^*)$. Theorem 5.2 shows that the ERM $\hat{f}_n$ reaches this minimal risk in the limit under suitable conditions. In the case of the square loss, the excess risk could be expressed as a distance between f and f^*, which made the results much stronger. For this general loss function this need not be the case, this makes the interpretation of Theorem 5.2 slightly less strong than in Theorem 4.2.

The original goal of regression was to estimate f_0. For the square loss it was known that $f^* = f_0 + \mu_\varepsilon$. For our general Lipschitz loss function, we know by Lemma 2.9 that $f_0 = f^*$ for all loss functions, excluding the quantile loss, introduced in Section 2.6, whenever ε has a symmetric density with zero mean. Hence the ERM is a consistent estimator for f_0 under symmetric, zero mean outliers, such as those coming from a t-distribution. In general, for our general Lipschitz continuous loss function, the relation between f_0 and f^* remains unknown for now, which makes it difficult to say how much better using a different loss than the square one, really is, when non-symmetric outliers are present.

5.2 Proof of Lemma 5.1.

In this section we present the proof of Lemma 5.1. The original proof was given by Shen et al. [2021] in Lemma 3.1. It should be noted that the proof is very similar to the proof of Lemma 4.1.

Let $S = \{(X_i, Y_i)\}_{i=1}^n$ be a sample of i.i.d. observations with distribution $Z := (X, Y)$. Also let $S' = \{(X_i', Y_i')\}_{i=1}^n$ be another sample independent of S and denote $Z_i' := (X_i', Y_i')$ and $Z_i := (X_i, Y_i)$. Define

$$g(f, Z_i) := L(f(X_i), Y_i) - L(f^*(X_i), Y_i)$$

for any f and observation Z_i. Note that the ERM $\hat{f}_n$ depends on the sample S, and its excess risk is

$$\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) = \mathbb{E}_Z \left[g(\hat{f}_n, Z) \right] = \mathbb{E}_{S'} \left[\frac{1}{n} \sum_{i=1}^n g(\hat{f}_n, Z_i') \right].$$

Hence its prediction error equals

$$\mathbb{E} \left\{ \mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right\} = \mathbb{E}_S \mathbb{E}_{S'} \left[\frac{1}{n} \sum_{i=1}^n g(\hat{f}_n, Z_i') \right]. \tag{5.1}$$

Define the *best in class estimator* f_φ^* as the estimator in the function class $\mathcal{NN}_p^{\mathcal{B}}$ with minimal L risk:

$$f_\varphi^* = \arg \min_{f \in \mathcal{NN}_p^{\mathcal{B}}} \mathcal{R}(f).$$

By the definition of the ERM, we have

$$\begin{aligned}
\mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n g(\hat{f}_n, Z_i) \right] &= \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n L(\hat{f}_n, Z_i) \right] - \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n L(f^*, Z_i) \right] \\
&\leq \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n L(f_\varphi^*, Z_i) \right] - \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n L(f^*, Z_i) \right] = \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n g(f_\varphi^*, Z_i) \right]. \quad (5.2)
\end{aligned}$$

Multiplying both sides of (5.2) by 2 and adding (5.1) gives

$$2\mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n g(\hat{f}_n, Z_i) \right] + \mathbb{E}_S \left\{ \mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right\} \leq 2\mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n g(f_\varphi^*, Z_i) \right] + \mathbb{E}_S \mathbb{E}_{S'} \left[\frac{1}{n} \sum_{i=1}^n g(\hat{f}_n, Z_i') \right],$$

so

$$\begin{aligned}
\mathbb{E}_S \left\{ \mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right\} &\leq \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}_{S'}[g(\hat{f}_n, Z_i')] - 2g(\hat{f}_n, Z_i) \right\} \right] + 2\mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n g(f_\varphi^*, Z_i) \right] \\
&= \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}_{S'}[g(\hat{f}_n, Z_i')] - 2g(\hat{f}_n, Z_i) \right\} \right] + 2 \left[\frac{1}{n} \sum_{i=1}^n \mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*) \right] \\
&= \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}_{S'}[g(\hat{f}_n, Z_i')] - 2g(\hat{f}_n, Z_i) \right\} \right] + 2 \left\{ \mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*) \right\}. \quad (5.3)
\end{aligned}$$

It is seen that the prediction error is upper bounded by the sum of an expectation of a stochastic term and an approximation error.

Next, we will focus on giving an upper bound of the first term on right-hand side of (5.3), and handle it with truncation. In the following, for ease of presentation, write

$$G(f, Z_i) := \mathbb{E}_{S'}[g(f, Z_i')] - 2g(f, Z_i) \text{ for any } f \in \mathcal{NN}_p^{\mathcal{B}}.$$

Observe that (5.3) can be written as

$$\mathbb{E}_S\left\{\mathcal{R}(\hat{f}_n)-\mathcal{R}(f^*)\right\} \le \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} G(\hat{f}_n, Z_i)\right] + 2\left\{\mathcal{R}(f^*_\varphi)-\mathcal{R}(f^*)\right\}. \tag{5.4}$$

Given a δ-uniform covering of $\mathcal{NN}_p^{\mathcal{B}}$, we denote the centers of the balls by f_j for $j=1,\dots,\mathcal{N}_{2n}$, where $\mathcal{N}_{2n} := \mathcal{N}_{2n}(\delta, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})$ is the uniform covering number with radius $\delta < \mathcal{B}$ under the norm $\|\cdot\|_\infty$ defined in (2.4). Notice the $2n$ because we want to cover with respect to both samples S and S' simultaneously. By definition of the covering number, there exists a random j^* such that $\|\hat{f}_n(X_i)-f_{j^*}(X_i)\|_\infty \le \delta$ and $\|\hat{f}_n(X_i')-f_{j^*}(X_i')\|_\infty \le \delta$ for any $(X_1,\dots,X_n,X_1',\dots,X_n') \in ([0,1]^d)^{2n}$. Hence for any $i=1,\dots,n$,

$$\begin{aligned}\left|g(\hat{f}_n, Z_i)-g(f_{j^*}, Z_i)\right| &= \left|L(\hat{f}_n(X_i), Y_i)-L(f_{j^*}(X_i), Y_i)\right| \\ &\le \lambda_L\left|\hat{f}_n(X_i)-f_{j^*}(X_i)\right| \le \lambda_L\delta,\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[g(\hat{f}_n, Z_i)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[g(f_{j^*}, Z_i)] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[g(\hat{f}_n, Z_i)-g(f_{j^*}, Z_i)] \\ &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[g(f_{j^*}, Z_i)] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S\left[\left|g(\hat{f}_n, Z_i)-g(f_{j^*}, Z_i)\right|\right] \\ &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[g(f_{j^*}, Z_i)] + \lambda_L\delta,\end{aligned}$$

where Z_i can be replaced by Z_i' since the covering covers with respect to both S and S'. Thus we have by Jensen's inequality,

$$\begin{aligned}\left|G(\hat{f}_n, Z_i)-G(f_{j^*}, Z_i)\right| &= \left|\mathbb{E}_{S'}[g(\hat{f}_n, Z_i')]-2g(\hat{f}_n, Z_i)-\mathbb{E}_{S'}[g(f_{j^*}, Z_i')]+2g(f_{j^*}, Z_i)\right| \\ &\le \left|\mathbb{E}_{S'}[g(\hat{f}_n, Z_i')]-\mathbb{E}_{S'}[g(f_{j^*}, Z_i')]\right| + 2\left|g(f_{j^*}, Z_i)-g(\hat{f}_n, Z_i)\right| \\ &\le \mathbb{E}_{S'}\left[\left|g(\hat{f}_n, Z_i')-g(f_{j^*}, Z_i')\right|\right] + 2\left|g(f_{j^*}, Z_i)-g(\hat{f}_n, Z_i)\right| \le 3\lambda_L\delta,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}G(\hat{f}_n, Z_i)\right] &= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}G(f_{j^*}, Z_i)\right] + \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}G(\hat{f}_n, Z_i)-G(f_{j^*}, Z_i)\right] \\ &\le \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}G(f_{j^*}, Z_i)\right] + \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}\left|G(\hat{f}_n, Z_i)-G(f_{j^*}, Z_i)\right|\right] \\ &\le \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}G(f_{j^*}, Z_i)\right] + 3\lambda_L\delta. \qquad (5.5)\end{aligned}$$

Let $\beta_n \ge \mathcal{B} \ge 1$ be a positive number that may depend on the sample size n. Denote T_{β_n} as the *truncation operator* at level β_n, that is,

$$T_{\beta_n}x = \begin{cases} x & \text{if } |x| \le \beta_n \\ \beta_n \cdot \operatorname{sign}(x) & \text{if } |x| > \beta_n \end{cases} \quad \text{for all } x \in \mathbb{R}.$$

Define the function $f^*_{\beta_n} : [0,1]^d \to \mathbb{R}$ pointwise by

$$f^*_{\beta_n}(x) = \arg\min_{f(x):\|f\|_\infty \le \beta_n} \mathbb{E}\left[L(f(X), T_{\beta_n}Y) \,\middle|\, X=x\right] \text{ for any } x \in \mathbb{R},$$

where the minimum is taken over all measurable functions $\mathcal{M}([0,1]^d)$. By definition of $f^*_{\beta_n}$ and f^*, we have for any measurable f,

$$\mathbb{E}[L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i)] \leq \mathbb{E}[L(f(X_i), T_{\beta_n}Y_i)] \quad \text{and} \quad \mathbb{E}[L(f^*(X_i), Y_i)] \leq \mathbb{E}[L(f(X_i), Y_i)].$$

For any $f \in \mathcal{NN}_p^{\mathcal{B}}$ and $i = 1, \ldots, n$, set $g_{\beta_n}(f, Z_i) := L(f(X_i), T_{\beta_n}Y_i) - L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i)$. Then we have

$$\begin{aligned}
\mathbb{E}[g(f, Z_i)] &= \mathbb{E}\left[L(f(X_i), Y_i) - L(f^*(X_i), Y_i)\right] \\
&= \mathbb{E}[g_{\beta_n}(f, Z_i)] + \mathbb{E}\left[L(f(X_i), Y_i) - L(f^*(X_i), Y_i)\right] \\
&\qquad + \mathbb{E}\left[L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i) - L(f(X_i), T_{\beta_n}Y_i)\right] \\
&= \mathbb{E}[g_{\beta_n}(f, Z_i)] + \mathbb{E}\left[L(f(X_i), Y_i) - L(f(X_i), T_{\beta_n}Y_i)\right] \\
&\qquad + \mathbb{E}\left[L(f^*(X_i), T_{\beta_n}Y_i) - L(f^*(X_i), Y_i)\right] \\
&\qquad + \mathbb{E}\left[L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i) - L(f^*(X_i), T_{\beta_n}Y_i)\right] \\
&\leq \mathbb{E}[g_{\beta_n}(f, Z_i)] + \mathbb{E}\left[L(f(X_i), Y_i) - L(f(X_i), T_{\beta_n}Y_i)\right] \\
&\qquad + \mathbb{E}\left[L(f^*(X_i), T_{\beta_n}Y_i) - L(f^*(X_i), Y_i)\right] \\
&\leq \mathbb{E}[g_{\beta_n}(f, Z_i)] + 2\lambda_L \mathbb{E}\left[\left|T_{\beta_n}Y_i - Y_i\right|\right] \leq \mathbb{E}[g_{\beta_n}(f, Z_i)] + 2\lambda_L \mathbb{E}\left[|Y_i| \mathbb{1}_{\{|Y_i| > \beta_n\}}\right] \\
&\leq \mathbb{E}[g_{\beta_n}(f, Z_i)] + 2\lambda_L \mathbb{E}\left[|Y_i| \frac{|Y_i|^{p-1}}{\beta_n^{p-1}}\right] \leq \mathbb{E}[g_{\beta_n}(f, Z_i)] + \frac{2\lambda_L}{\beta_n^{p-1}} \mathbb{E}|Y_i|^p.
\end{aligned}$$

See the proof of Lemma 4.1 for more details as to why $\mathbb{E}\left[\left|T_{\beta_n}Y_i - Y_i\right|\right] \leq \mathbb{E}\left[|Y_i| \mathbb{1}_{\{|Y_i| > \beta_n\}}\right]$.

By assumption, $\mathbb{E}|\varepsilon_i|^p < \infty$, hence $\mathbb{E}|Y_i|^p < \infty$ since $f_0(X_i)$ is a bounded random variable. Similarly,

$$\begin{aligned}
\mathbb{E}[g_{\beta_n}(f, Z_i)] &= \mathbb{E}[g(f, Z_i)] + \mathbb{E}[L(f^*(X_i), Y_i) - L(f^*_{\beta_n}(X_i), Y_i)] \\
&\qquad + \mathbb{E}[L(f(X_i), T_{\beta_n}Y_i) - L(f(X_i), Y_i)] \\
&\qquad + \mathbb{E}[L(f^*_{\beta_n}(X_i), Y_i) - L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i)] \\
&\leq \mathbb{E}[g(f, Z_i)] + \mathbb{E}[L(f(X_i), T_{\beta_n}Y_i) - L(f(X_i), Y_i)] \\
&\qquad + \mathbb{E}[L(f^*_{\beta_n}(X_i), Y_i) - L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i)] \\
&\leq \mathbb{E}[g(f, Z_i)] + \frac{2\lambda_L}{\beta_n^{p-1}} \mathbb{E}|Y_i|^p.
\end{aligned}$$

Note that above inequalities also hold for $g(f, Z_i')$ and $g_{\beta_n}(f, Z_i)$. Define

$$G_{\beta_n}(f, Z_i) := \mathbb{E}_{S'}[g_{\beta_n}(f, Z_i')] - 2g_{\beta_n}(f, Z_i)$$

for any $f \in \mathcal{NN}_p^{\mathcal{B}}$, then

$$\begin{aligned}
\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G(f_{j^*}, Z_i)\right] &= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \{\mathbb{E}_{S'}[g(f_{j^*}, Z_i')] - 2g(f_{j^*}, Z_i)\}\right] \\
&= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \{\mathbb{E}_{S'}[g(f_{j^*}, Z_i')] - 2g(f_{j^*}, Z_i)\}\right] + \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i)\right] \\
&\quad - \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \{\mathbb{E}_{S'}[g_{\beta_n}(f_{j^*}, Z_i')] - 2g_{\beta_n}(f_{j^*}, Z_i)\}\right] \\
&= \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i)\right] \\
&\quad + \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n \{\mathbb{E}_{S'}[g(f_{j^*}, Z_i') - g_{\beta_n}(f_{j^*}, Z_i')] + 2(g_{\beta_n}(f_{j^*}, Z_i) - g(f_{j^*}, Z_i))\}\right] \\
&\leq \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i)\right] + \frac{6\lambda_L}{\beta_n^{p-1}}\mathbb{E}|Y_i|^p. \qquad (5.6)
\end{aligned}$$

Observe that for any $f \in \mathcal{NN}_p^{\mathcal{B}}$, we have

$$\begin{aligned}
|g_{\beta_n}(f, Z_i)| &= |L(f(X_i), T_{\beta_n}Y_i) - L(f^*_{\beta_n}(X_i), T_{\beta_n}Y_i)| \\
&\leq \lambda_L|f(X_i) - f^*_{\beta_n}(X_i)| \leq \lambda_L(|f(X_i)| + |f^*_{\beta_n}(X_i)|) \\
&\leq 2\lambda_L\beta_n \leq 4\lambda_L\beta_n,
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sigma_g^2(f) := \mathrm{Var}(g_{\beta_n}(f, Z_i)) &\leq \mathbb{E}[g_{\beta_n}(f, Z_i)^2] \leq \mathbb{E}[|g_{\beta_n}(f, Z_i)|g_{\beta_n}(f, Z_i)] \\
&\leq 4\lambda_L\beta_n\mathbb{E}[g_{\beta_n}(f, Z_i)],
\end{aligned}$$

where we used the fact that $g_{\beta_n}(f, Z_i) \geq 0$. For each f_j and any $t > 0$, let $u := \frac{t}{2} + \frac{\sigma_g^2(f_j)}{8\lambda_L\beta_n}$, by

Bernstein's inequality [Boucheron et al., 2013, section 2.7],

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_j, Z_i) > t\right) &= \mathbb{P}\left(\mathbb{E}_{S'}[g_{\beta_n}(f_j, Z_1')] - \frac{2}{n}\sum_{i=1}^{n} g_{\beta_n}(f_j, Z_i) > t\right) \\
&= \mathbb{P}\left(\mathbb{E}_{S'}[g_{\beta_n}(f_j, Z_1')] - \frac{1}{n}\sum_{i=1}^{n} g_{\beta_n}(f_j, Z_i) > \frac{t}{2} + \frac{1}{2}\mathbb{E}_{S'}[g_{\beta_n}(f_j, Z_i')]\right) \\
&\leq \mathbb{P}\left(\mathbb{E}_{S'}[g_{\beta_n}(f_j, Z_1')] - \frac{1}{n}\sum_{i=1}^{n} g_{\beta_n}(f_j, Z_i) > \frac{t}{2} + \frac{1}{2}\frac{\sigma_g^2(f_j)}{4\lambda_L\beta_n}\right) \\
&= \mathbb{P}\left(n\mathbb{E}_{S'}[g_{\beta_n}(f_j, Z_1')] - \sum_{i=1}^{n} g_{\beta_n}(f_j, Z_i) > nu\right) \\
&\leq \exp\left(-\frac{n^2u^2}{2\left(n\sigma_g^2(f_j) + \frac{4nu\lambda_L\beta_n}{3}\right)}\right) \\
&= \exp\left(-\frac{nu^2}{(2u-t)8\lambda_L\beta_n + \frac{8u\lambda_L\beta_n}{3}}\right) \\
&\leq \exp\left(-\frac{nu^2}{16u\lambda_L\beta_n + \frac{16u\lambda_L\beta_n}{3}}\right) \\
&= \exp\left(-\frac{1}{16+\frac{16}{3}}\cdot\frac{nu}{\lambda_L\beta_n}\right) \leq \exp\left(-\frac{1}{16+\frac{16}{3}}\cdot\frac{n\frac{1}{2}t}{\lambda_L\beta_n}\right) \\
&\leq \exp\left(-\frac{1}{43}\cdot\frac{nt}{\lambda_L\beta_n}\right).
\end{aligned}$$

Using the union bound, above inequality leads to a tail probability bound of $\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)$, that is

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i) > t\right) &\leq \mathbb{P}\left(\bigcup_{j=1}^{\mathcal{N}_{2n}}\left\{\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_j, Z_i) > t\right\}\right) \\
&\leq \sum_{j=1}^{\mathcal{N}_{2n}} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_j, Z_i) > t\right) \\
&\leq \mathcal{N}_{2n}\exp\left(-\frac{1}{43}\cdot\frac{nt}{\lambda_L\beta_n}\right) \leq 2\mathcal{N}_{2n}\exp\left(-\frac{1}{43}\cdot\frac{nt}{\lambda_L\beta_n}\right).
\end{aligned}$$

In the same way as we have done in the proof of Lemma 4.1, we obtain for any $a_n > 0$,

$$\begin{aligned}
\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)\right] &\leq a_n + \int_{a_n}^{\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i) > t\right) dt \\
&\leq a_n + \int_{a_n}^{\infty} 2\mathcal{N}_{2n}\exp\left(-\frac{1}{43}\cdot\frac{nt}{\lambda_L\beta_n}\right) dt \\
&= a_n + 2\mathcal{N}_{2n}\exp\left(-a_n\cdot\frac{n}{43\lambda_L\beta_n}\right)\frac{43\lambda_L\beta_n}{n}.
\end{aligned}$$

Choosing $a_n := \log(2\mathcal{N}_{2n})\cdot\frac{43\lambda_L\beta_n}{n}$, we have

$$\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)\right] \leq \frac{43\lambda_L\beta_n\left(1 + \log(2\mathcal{N}_{2n})\right)}{n}. \tag{5.7}$$

Setting $\delta = \frac{1}{n}$ and $\beta_n = \max(\mathcal{B}, n^{\frac{1}{p}})$ and combining (5.4), (5.5), (5.6) and (5.7), we get

$$\begin{aligned}
\mathbb{E}_S\left\{\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right\} &\leq \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G(\hat{f}_n, Z_i)\right] + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} \\
&\leq \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G(f_{j^*}, Z_i)\right] + 3\lambda_L\frac{1}{n} + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} \\
&\leq \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i)\right] + \frac{6\lambda_L}{\beta_n^{p-1}}\mathbb{E}|Y_i|^p + 3\lambda_L\frac{1}{n} + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} \\
&\leq \frac{43\lambda_L\beta_n}{n}\left(\log(2\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + \frac{6\lambda_L}{\beta_n^{p-1}}\mathbb{E}|Y_i|^p + \frac{3\lambda_L}{n} \\
&\quad + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} \\
&= \frac{\lambda_L}{n}\left(43\beta_n\left(\log(2\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + \frac{6\mathbb{E}|Y_i|^p}{\beta_n^{p-1}}\cdot n + 3\right) \\
&\quad + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} = (*)
\end{aligned}$$

Observe that $\beta_n = \max(\mathcal{B}, n^{\frac{1}{p}}) \leq \mathcal{B}n^{\frac{1}{p}}$ and $\beta_n \geq n^{\frac{1}{p}}$. Hence

$$\begin{aligned}
(*) &\leq \frac{\lambda_L}{n}\left(43n^{\frac{1}{p}}\mathcal{B}\left(\log(2\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}})) + 1\right) + 6\mathbb{E}|Y_i|^p n^{\frac{1}{p}} + 3\right) + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} \\
&= \frac{\lambda_L}{n}\left(n^{\frac{1}{p}}\left[43\mathcal{B}\left(\log\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) + \log 2 + 1\right) + 6\mathbb{E}|Y_i|^p\right] + 3\right) + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\}.
\end{aligned}$$

Similarly to the proof of Lemma 4.1, it follows that

$$\begin{aligned}
(*) &\leq \frac{\lambda_L}{n}\left(c_1 n^{\frac{1}{p}}\mathcal{B}\log\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) + 3\right) + 2\left\{\mathcal{R}(f_\varphi^*) - \mathcal{R}(f^*)\right\} \\
&\leq c_0\left(\frac{\lambda_L\mathcal{B}}{n^{1-\frac{1}{p}}}\right)\log\mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\left\{\mathcal{R}(f) - \mathcal{R}(f^*)\right\},
\end{aligned}$$

where $c_0 > 0$ is a constant independent of $n, d, \lambda_L, \mathcal{B}, \mathcal{S}, \mathcal{W}$ and $\mathcal{D}$. This proves the first result.

From the proof of Lemma 4.1, it is known that

$$\mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{NN}_p^{\mathcal{B}}) \leq C\mathcal{SD}\log(\mathcal{S})\log n,$$

for some constant $C > 0$. Therefore it follows that,

$$\mathbb{E}_S\left\{\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*)\right\} \leq C_0\left(\frac{\log n}{n^{1-\frac{1}{p}}}\right)\lambda_L\mathcal{BDS}\log\mathcal{S} + 2\inf_{f\in\mathcal{NN}_p^{\mathcal{B}}}\left\{\mathcal{R}(f) - \mathcal{R}(f^*)\right\},$$

where $C_0 := C\cdot c_0$ is a constant independent of $n, d, \lambda_L, \mathcal{B}, \mathcal{S}, \mathcal{W}$ and $\mathcal{D}$. □

6 Experiments.

In this section we first introduce the procedure of estimating the empirical risk minimizer from some function class $\mathcal{NN}_p^{\mathcal{B}}$. Then, we demonstrate the approximation of the empirical risk minimizer for some simple, univariate regression functions. This is done under the contamination model for ε with a few different outlier densities. Finally, the prediction error is estimated for different sample sizes, and compared to the theoretical rate provided by Lemma 4.1 and Lemma 5.1.

6.1 Estimating and evaluating.

In for instance, linear regression, one has an explicit solution of the parameters of the ERM as function of the random sample S. For neural networks such an explicit formulation does not exists. Instead, optimization algorithms have been developed based on gradient descent. We will be using the Adam optimization algorithm [Kingma and Ba, 2017]. Throughout our simulations the TensorFlow library [Abadi et al., 2015] in Python is used, with the standard implemented hyperparameters for the Adam optimizer.

Consider a random sample S. The Adam optimizer randomly initializes the weights and biases of the network and optimizes these parameters using a gradient descent based method. By this process, an estimator $\hat{f}_{\text{train}} \in \mathcal{NN}_p^{\mathcal{B}}$ is obtained. In the training process, the empirical risk $\mathcal{R}_n$ is optimized for, hence $\hat{f}_{\text{train}}$ will hopefully be a good approximation of the true ERM $\hat{f}_n$. Throughout our experiments, we train each network for 1500 epochs with a batch size of $\frac{n}{15}$. For all other parameters, the standard implemented values are used. Early stopping with a patience of 50 is also used to prevent unnecessary compute time.

Notice that the estimator $\hat{f}_{\text{train}}$ depends randomly on the initialization of the parameters. Hence one can have a bad, or good initialization. Since the optimizer can get stuck in local minima, this is an issue. Therefore the network is trained $t \in \mathbb{N}$ times on the same sample with different parameter initialization, resulting in estimators $\hat{f}_{\text{train},1}, \dots, \hat{f}_{\text{train},t}$, $t \in \mathbb{N}$. From these estimators the estimator with the least empirical risk is chosen, denoted by $\hat{f}^t_{\text{train}}$. That is,

$$\hat{f}^t_{\text{train}} = \arg \min_{1 \leq i \leq t} \mathcal{R}_n(\hat{f}_{\text{train},i}).$$

The hope is that by training multiple networks and taking the one with minimal empirical risk, one gets a better approximation of the ERM $\hat{f}_n$.

For any estimator f, possibly depending the sample S, the risk $\mathcal{R}(f)$ is estimated using Monte Carlo estimation. That is, we generate $(X_1, Y_1), \dots, (X_N, Y_N)$ with distribution $Z = (X, Y)$ independent of the sample S. Then, by the law of large numbers,

$$\hat{\mathcal{R}}(f) := \frac{1}{N} \sum_{i=1}^{N} L(f(X), Y) \xrightarrow{p} \mathbb{E}_Z \left[L(f(X), Y) \right] = \mathcal{R}(f).$$

If the second moments of $\mathcal{R}(f)$ are finite, one can also obtain 95% confidence intervals for the true risk $\mathcal{R}(f)$ using the central limit theorem.

The empirical risk $\mathcal{R}_n$ and risk $\mathcal{R}$ depend on a chosen loss function L. In our simulations, we consider the square loss, the Huber loss with parameter $\zeta = 1$ and Tukey's biweight loss with parameter $t = 4.685$ [Belagiannis et al., 2015].

6.2 Fitting univariate functions.

In the subsection we estimate various regression functions f_0 for the square, Huber and Tukey loss. While using the contamination model from Section 2.6 with a mixture rate of $\alpha = 0.2$, standard observation distribution $\mathrm{N}(0, 0.02)$ and three different outlier densities; normally distributed $\mathrm{N}(0, 0.2)$, Fréchet distributed with $\lambda = 3$ and t-distribution with $\nu = 3$. The distribution used for X is a uniform distribution $\mathrm{U}(0, 1)$.

The regression functions to be estimated are

$$f_0(x) = x,$$

$$f_0(x) = x^2,$$

$$f_0(x) = \frac{1}{2} \left(1 + \sin(6\pi x)\right),$$

for any $x \in [0,1]$. Observe that above functions have range $[0,1]$.

Throughout this subsection, the network architecture will be fixed at

$$p = (1, 50, 100, 200, 200, 200, 100, 50, 1).$$

For each outlier distribution, a sample S is generated with sample size $n = 400$. Then, for a given loss function, $\hat{f}_{\text{train},1}, \dots, \hat{f}_{\text{train},30}$ are obtained by training the networks. Again, the estimator $\hat{f}^{30}_{\text{train}}$ with the minimum empirical risk is used as our approximation for the ERM $\hat{f}_n$, that is,

$$\hat{f}^{30}_{\text{train}} = \arg\min_{1 \le i \le 30} \mathcal{R}_n(\hat{f}_{\text{train},i}).$$

In Fig. 6.1, the samples and ERM estimates are displayed for the regression function $f_0(x) = x$. For the normal outliers, which are relatively well behaved, all loss functions perform very well for such a simple regression function. For the t-distributed outliers, we know that the ERM is a consistent estimator for f_0, which we see in Fig. 6.1, though some error is made due to the noise. Finally, for the Fréchet outliers, it can be clearly seen that all loss functions result in a biased ERM. For the square loss it was already known that a bias is created when non-zero mean outliers are present. Do note that the bias of the Huber and Tukey loss is slightly smaller than the bias of the square loss.

In Fig. 6.2, the same conclusions can be made as from Fig. 6.1. Though one should note that, with the Fréchet outliers, the larger bias of the square loss compared to the Tukey and Huber loss is even clearer.

Finally, in Fig. 6.3, a more complicated regression function has been used. In the case of well behaved outliers, that is, $\mathrm{N}(0, 0.2)$ outliers, all loss functions seem to result in a good approximation of the regression function. When more extreme outliers are introduced, as is the case with $t(3)$ and Fréchet(3) outliers, the estimation algorithm itself seems to get unstable. This can by seen by the fact that the square loss results in an almost straight line, which is clearly not equal to the empirical risk minimizer. This highlights that even though the ERM has nice theoretical properties, it can be difficult to well approximate the ERM even in quite simple circumstances.

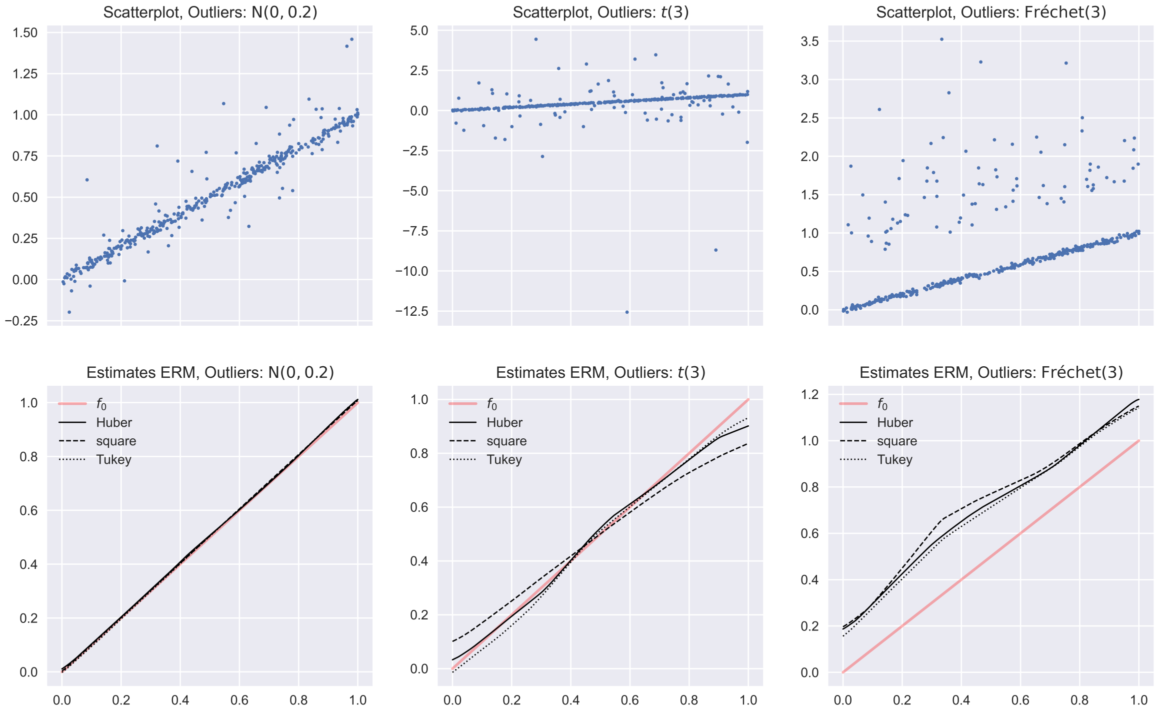


Figure 6.1: On the top row, three scatterplots are shown. Each sample, with sample size $n = 400$, has been generated with the relation $Y_i = f_0(X_i) + \varepsilon_i$, where $X_i \sim \mathrm{U}(0,1)$, $f_0(x) = x$ and ε_i a mixture between $\mathrm{N}(0, 0.02)$ and the given outlier distribution with mixture rate $\alpha = 0.2$. Below, for three loss functions, the ERM has been estimated using the sample above it and displayed in black, along with the true regression function f_0 in red.

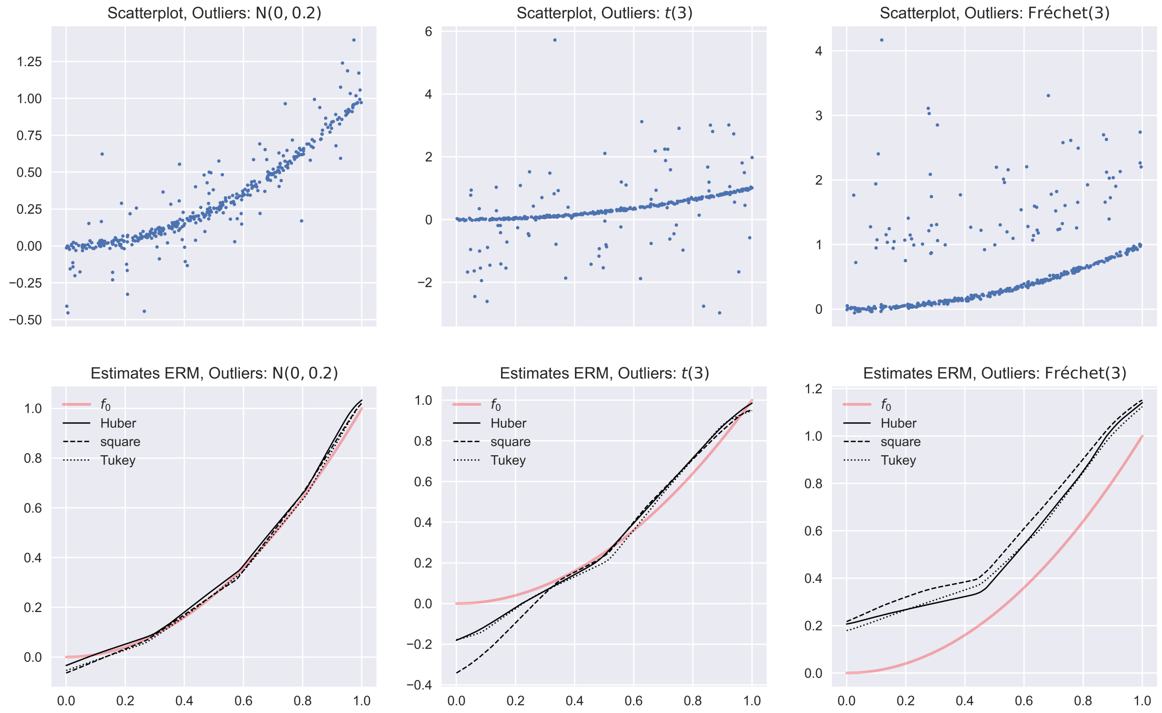


Figure 6.2: On the top row, three scatterplots are shown. Each sample, with sample size $n = 400$, has been generated with the relation $Y_i = f_0(X_i) + \varepsilon_i$, where $X_i \sim \mathrm{U}(0, 1)$, $f_0(x) = x^2$ and ε_i a mixture between $\mathrm{N}(0, 0.02)$ and the given outlier distribution with mixture rate $\alpha = 0.2$. Below, for three loss functions, the ERM has been estimated using the sample above it and displayed in black, along with the true regression function f_0 in red.

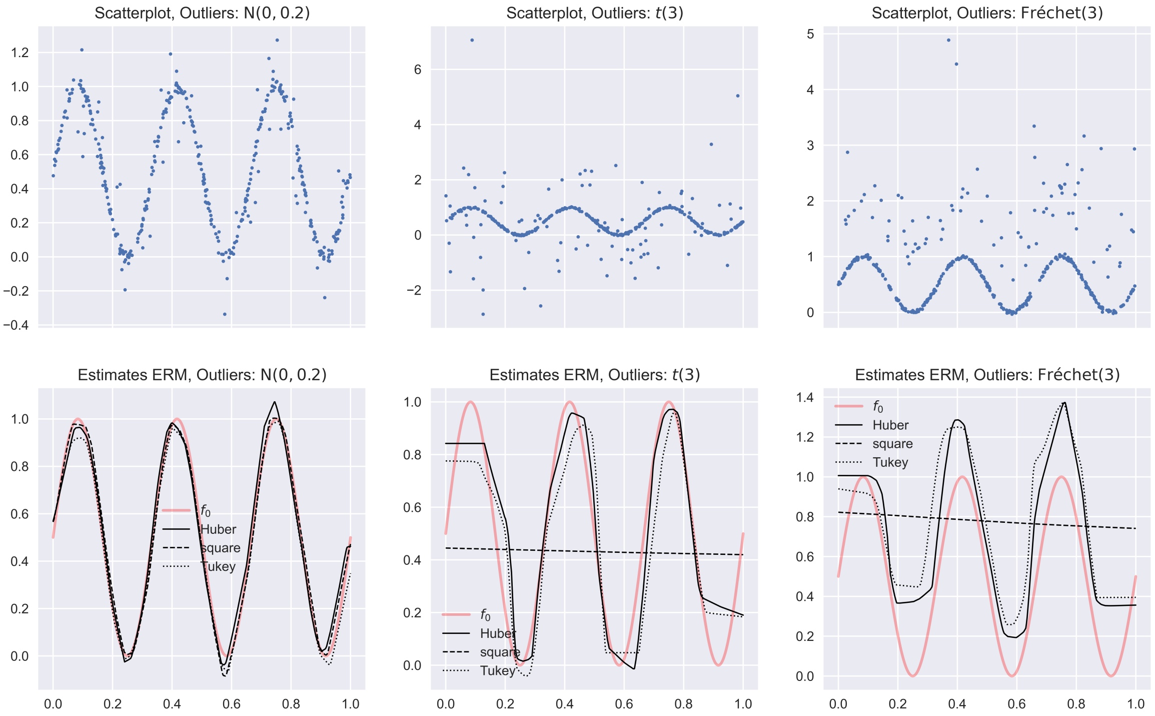


Figure 6.3: On the top row, three scatterplots are shown. Each sample, with sample size $n = 400$, has been generated with the relation $Y_i = f_0(X_i) + \varepsilon_i$, where $X_i \sim \mathrm{U}(0,1)$, $f_0(x) = \frac{1}{2}(1 + \sin(6\pi x))$ and ε_i a mixture between $\mathrm{N}(0, 0.02)$ and the given outlier distribution with mixture rate $\alpha = 0.2$. Below, for three loss functions, the ERM has been estimated using the sample above it and displayed in black, along with the true regression function f_0 in red.

6.3 Checking the convergence rate.

Consider the regression model in (1.1). If the ERM $\hat{f}_n$ is used to estimate f^*, Lemma 4.1 and Lemma 5.1 give us a convergence rate depending on the moments of ε if it is assumed that $f^* \in \mathcal{NN}_p^{\mathcal{B}}$. In the following we estimate the prediction error for the regression function $f_0(x) = x$ with various outlier densities. Then, the convergence rate on the estimated prediction errors can be compared with the theoretical ones. This helps us either indicate that the theoretical rate is sharp, or provide an indication that the theoretical rate can be further improved.

In order to estimate the prediction error $\mathbb{E}_S \left\{ \mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right\}$, one needs to know the function f^*. Under the square loss we have seen in Lemma 2.7 that $f^* = f_0 + \mu_\varepsilon$, but for the Lipschitz continuous loss function such a relation is not generally known. Instead of estimating the prediction error directly, we will estimate the *adjusted prediction error* $\mathbb{E}_S \left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f_0) \right]$. It is related to the prediction error by

$$\mathbb{E}_S \left[\mathcal{R}(\hat{f}_n) - \mathcal{R}(f_0) \right] = \mathbb{E}_S \left\{ \mathcal{R}(\hat{f}_n) - \mathcal{R}(f^*) \right\} + \left\{ \mathcal{R}(f^*) - \mathcal{R}(f_0) \right\}.$$

For the square loss it follows by Lemma 2.7 that

$$\left\{ \mathcal{R}(f^*) - \mathcal{R}(f_0) \right\} = -\mu_\varepsilon^2,$$

hence the adjusted prediction error will coincide with the prediction error whenever $\mu_\varepsilon = 0$. When ε is symmetric with mean zero, and the other assumption on L from Lemma 2.9 are satisfied, we have $f^* = f_0$, which implies $\mathcal{R}(f^*) - \mathcal{R}(f_0) = 0$.

In our simulations, the contamination model from Section 2.6 is considered with mixture rate $\alpha = 0.05$. For the standard observation, a $\mathrm{N}(0, 0.02)$ distribution is used. For the outliers we consider three different distributions; $\mathrm{N}(0, 0.2)$, $t(3)$ and Fréchet(3). The network architecture will remain fixed with

$$p = (1, 5, 1),$$

which ensures $f^* \in \mathcal{NN}_p^{\mathcal{B}}$. For each combination of loss function and outlier density, Lemma 3.7 or Lemma 5.1 gives a convergence rate on the prediction error. These are given in Table 4. These theoretical rates can then be compared to an observed rate.

	$\mathrm{N}(0, 0.2)$	$t(3)$	Fréchet(3)
Square loss	$O\left(\frac{\log n}{n}\right)$	$O\left(\sqrt[5]{\frac{\log n}{n}}\right)$	$O\left(\sqrt[5]{\frac{\log n}{n}}\right)$
Huber loss	$O\left(\frac{\log n}{n}\right)$	$O\left(\frac{\log n}{\sqrt[3]{n}}\right)$	$O\left(\frac{\log n}{\sqrt[3]{n}}\right)$
Tukey loss	$O\left(\frac{\log n}{n}\right)$	$O\left(\frac{\log n}{\sqrt[3]{n}}\right)$	$O\left(\frac{\log n}{\sqrt[3]{n}}\right)$

Table 4: The convergence rate of the prediction error of the given loss function with outlier density when assuming $f^* \in \mathcal{NN}_p^{\mathcal{B}}$ and normally distributed standard observations. The rates follow from Lemma 4.1 and Lemma 5.1.

To obtain an empirical convergence rate, we must estimate the adjusted prediction error for various sample sizes n. The sample size start at $n = 50$ and gets doubled 9 times, thus

$$n \in \{50, 100, 200, 400, 800, 1600, 3200, 6400, 12800, 25600\}.$$

For each sample size n, we estimate the adjusted prediction error as described in Section 6.1 with $t = 5$ and $N = 100000$ for each loss function and outlier density. The result can be seen in Fig. 6.4 along with 95% confidence intervals. In Fig. 6.4(a) and (b), we have symmetric zero-mean outliers. Hence by Lemma 2.7 and Lemma 2.9 we know that $f^* = f_0$, which implies $\mathcal{R}(f^*) - \mathcal{R}(f_0) = 0$. Thus the adjusted prediction error agrees with the prediction error.

Note that since

$$\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n)-\mathcal{R}(f_0)\right]=\mathbb{E}_S\left\{\mathcal{R}(\hat{f}_n)-\mathcal{R}(f^*)\right\}+\{\mathcal{R}(f^*)-\mathcal{R}(f_0)\},$$

the adjusted prediction error will converge to $\{\mathcal{R}(f^*)-\mathcal{R}(f_0)\}$ because the prediction error converges to zero. Visually, we see in Fig. 6.4(a) and (b) that the adjusted prediction error converges to zero, which then must be equal to $\{\mathcal{R}(f^*)-\mathcal{R}(f_0)\}$.

For the square loss, and under suitable assumption that are satisfied, we know that the adjusted prediction error satisfies

$$\begin{aligned}\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n)-\mathcal{R}(f_0)\right]&=\mathbb{E}_S\left\{\mathcal{R}(\hat{f}_n)-\mathcal{R}(f^*)\right\}+\{\mathcal{R}(f^*)-\mathcal{R}(f_0)\}\\&\to\{\mathcal{R}(f^*)-\mathcal{R}(f_0)\}=-\mu_\varepsilon^2=-\alpha^2\mu_{\text{outl}}^2,\end{aligned}$$

where α is the mixture rate and μ_{outl} is the expectation of the outlier density. When Fréchet(3) outliers are present, $\mu_{\text{outl}}=\Gamma\left(\frac{2}{3}\right)$. Hence,

$$\mathbb{E}_S\left[\mathcal{R}(\hat{f}_n)-\mathcal{R}(f_0)\right]\to-\alpha^2\Gamma\left(\frac{2}{3}\right)^2\approx-0.00458\ldots.$$

This exactly agrees with the convergence of the square loss in Fig. 6.4(c). In Fig. 6.4(c) one also sees that $\mathcal{R}(f^*)-\mathcal{R}(f_0)\neq 0$ for the Huber and Tukey loss. It clear that this difference is less than for the square loss. In that sense, the Huber and Tukey loss are more robust against non-symmetric outliers than the square loss, which was the original motivation for introducing these other loss functions.

Now one would like to compare the observed rate with our theoretical rates in Table 4. If our observed rate matches the theoretical rate, dividing by the theoretical rate should result in a constant line (after first centering the observed rate such that is converges to 0). If the observed rate is faster than the theoretical rate, the line should go down.

In Fig. 6.5, the observed rates from Fig. 6.4 have been divided by the theoretical rates from Table 4. For N$(0, 0.2)$ outliers [see Fig. 6.5(a)], we observe that for small sample sizes, the observed rate is faster than the theoretical rate. Note however, that the theoretical rate only applies for sample sizes greater or equal to the pseudodimension, hence it is not a problem that the observed rate outperforms the theoretical rate for small sample sizes. For larger sample sizes we observe an almost straight line in Fig. 6.5(a), suggesting that our theoretical rate is sharp.

In both Fig. 6.5(b) and (c), we observe that the fraction keeps going down, even for larger samples sizes. This shows that when ε has only a small number of finite moments, the observed rate is faster than our theoretical rate, indicating that the theoretical rate can be sharpened further. One might wonder what the correct rate is in Fig. 6.5(b) and (c). In Fig. 6.6, all observed rates have been divided by $\frac{\log n}{n}$. Note especially that for the t and Fréchet outlier, the observed seems to equal $\frac{\log n}{n}$, instead of our predicted rate. However, this does not show the rate can be sharpened to $\frac{\log n}{n}$ in general. The correct theoretical rate needs to be further explored in future work.

It could be that our observed rate is only so fast because the network architecture is allowed to be fixed since the regression function lies inside the network function class. If one has a more complicated regression function, one has to grow the network with the sample size. It might be possible that this will slow down the observed rate to the rate we predict theoretically. Further simulations are needed to know for sure.

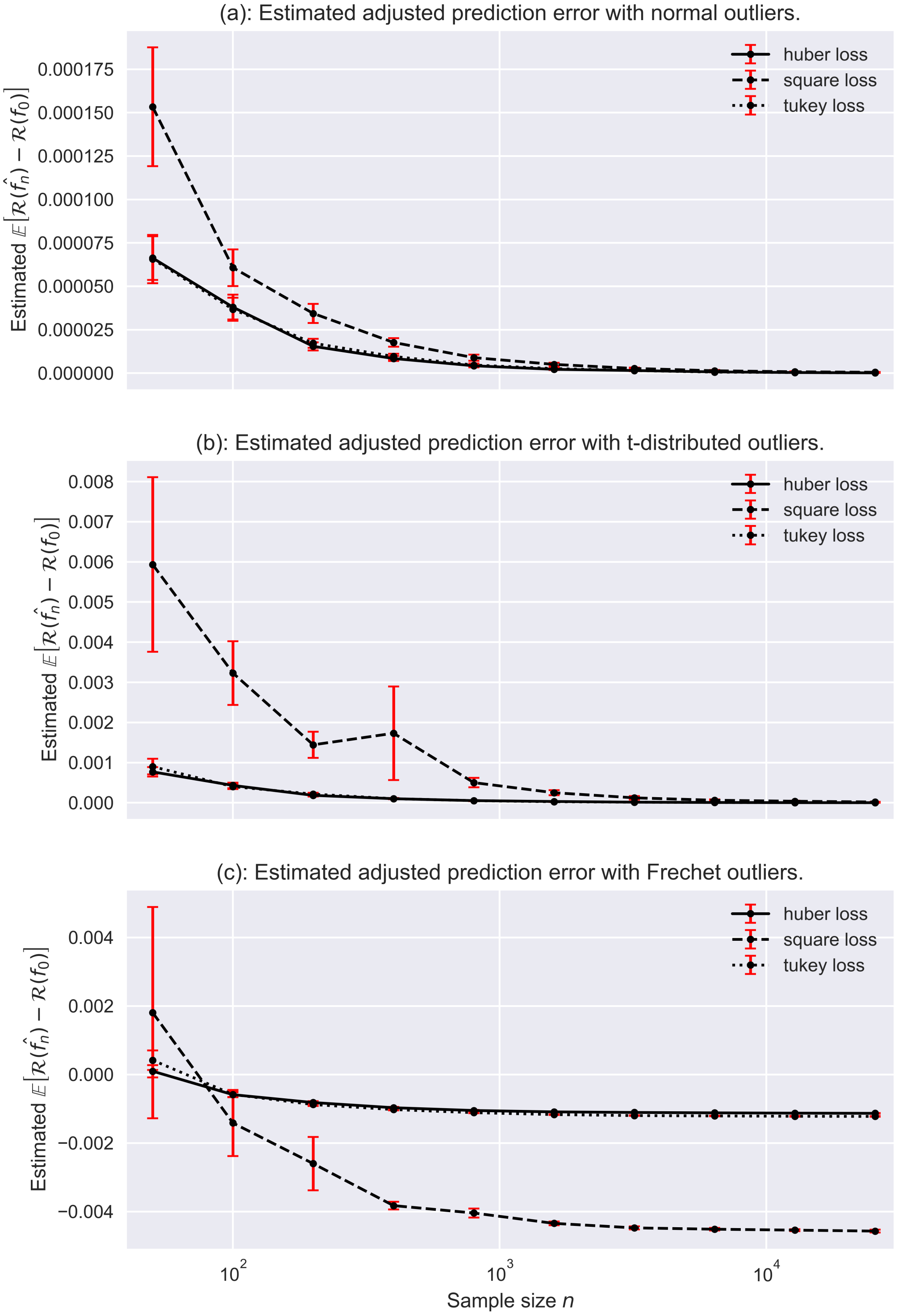


Figure 6.4: The adjusted prediction error with different outlier distributions. In each sub figure, the Huber loss, square loss and Tukey loss have been used. For each sample size, the estimated adjusted prediction error has been calculated along with an asymptotic 95% CI. The outliers densities used are $\mathrm{N}(0, 0.2)$, $t(3)$ and Fréchet(3).

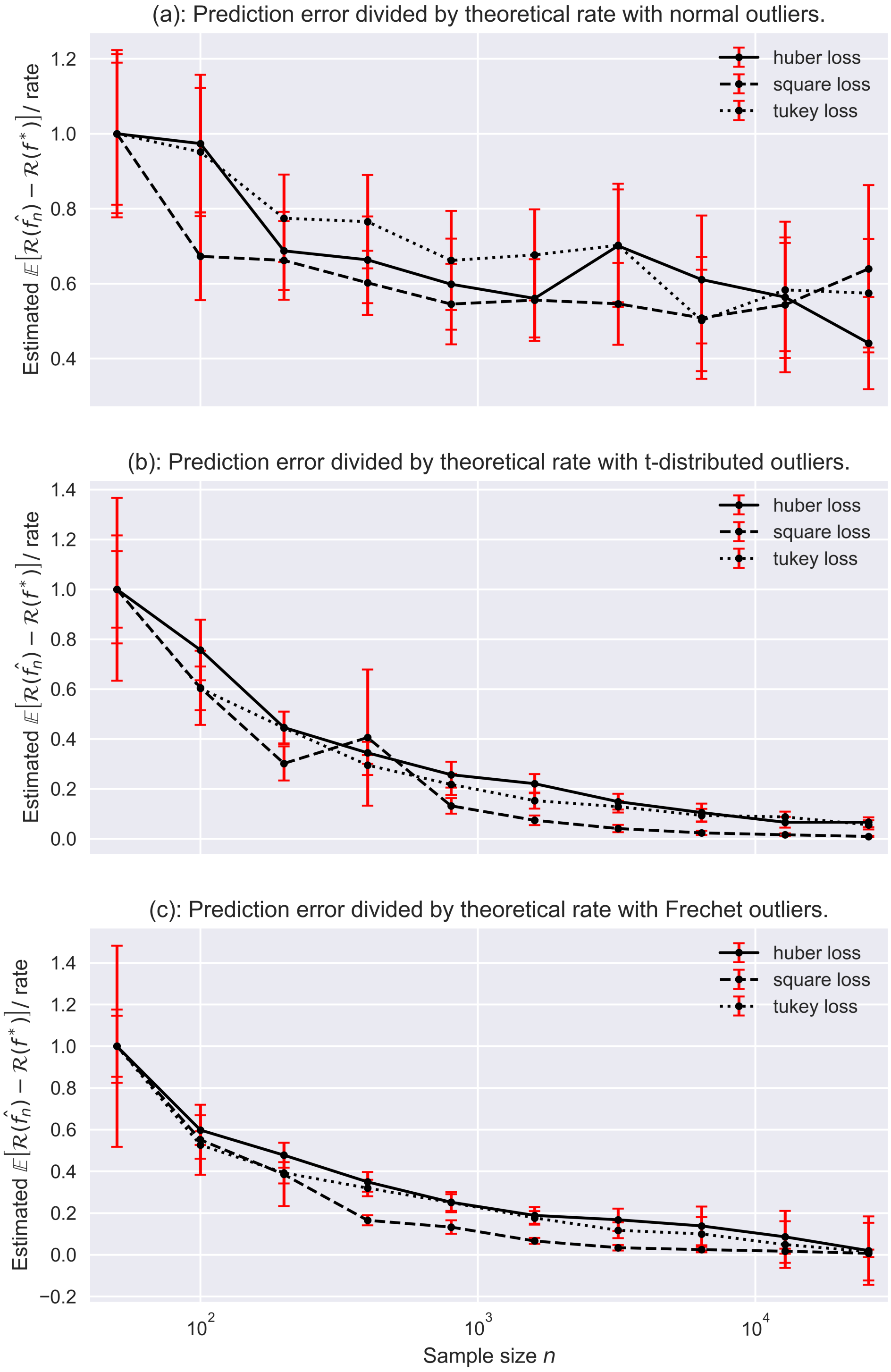
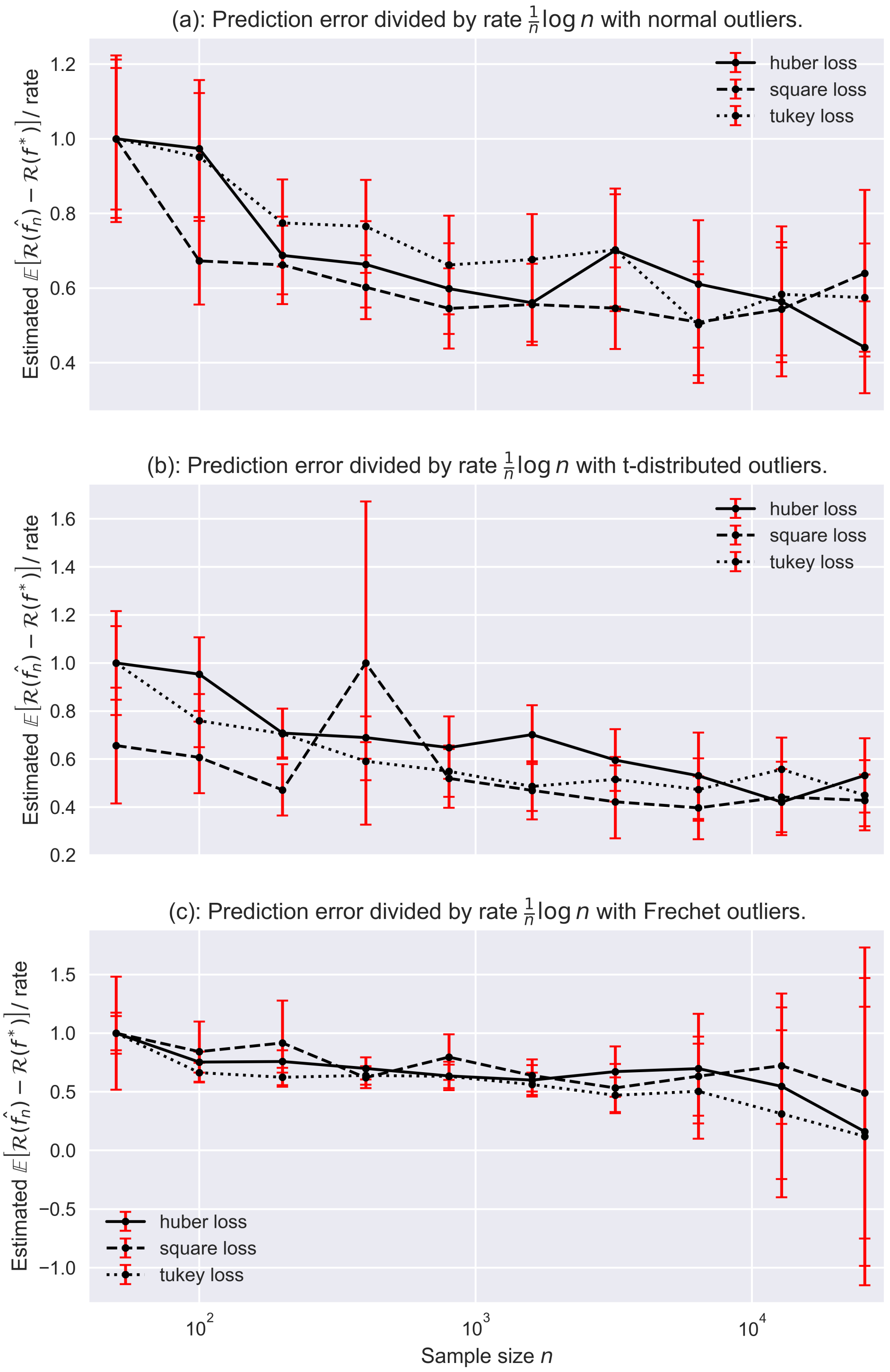


Figure 6.5: The prediction error divided by the theoretical convergence rates from Table 4. In each sub figure, the Huber loss, square loss and Tukey loss have been used. For each sample size, the estimated fraction has been calculated along with an asymptotic 95% CI. Furthermore, each line has been divided by a proper constant such that the results are of the same order. The outliers densities used are $\mathrm{N}(0, 0.2)$, $t(3)$ and Fréchet(3).


(a): Prediction error divided by rate $\frac{1}{n}\log n$ with normal outliers.

(b): Prediction error divided by rate $\frac{1}{n}\log n$ with t-distributed outliers.

(c): Prediction error divided by rate $\frac{1}{n}\log n$ with Frechet outliers.

Figure 6.6: The prediction error divided by the rate $\frac{\log n}{n}$. In each sub figure, the Huber loss, square loss and Tukey loss have been used. For each sample size, the estimated fraction has been calculated along with an asymptotic 95% CI. Furthermore, each line has been divided by a proper constant such that the results are of the same order. The outliers densities used are $\mathrm{N}(0, 0.2)$, $t(3)$ and Fréchet(3).

7 Conclusion and discussion.

We have seen non-asymptotic error bounds on the prediction error for the square loss, and for a more general Lipschitz continuous loss function. When the regression function belongs to the class of neural networks, convergence is obtained with a rate depending on the number of finite moments of ε.

Using these error bounds, we showed that the empirical risk minimizer is a consistent estimator for the unknown regression function f_0 when the error ε has zero mean and the square loss is used. For the Lipschitz continuous loss function, the density of ε also has to be symmetric. This shows that the ERM is robust against symmetric outliers generated by for example, a t-distribution. For the square loss it has been proven that $f^* = f_0 + \mu_\varepsilon$. This shows that when outliers are present that do not have mean zero, the ERM is a biased estimator for f_0. The general relation between f^* and f_0 remains unknown for our general Lipschitz continuous loss function. It is however crucial to understand this better in order to say something about the effect of non-zero mean outliers under a robust loss function.

Finally, we estimated the ERM for some basic univariate regression functions. It became clear that even for quite simple regression functions, the estimation procedure can be unstable. Then, we have seen empirically that our theoretical convergence rate look tight when ε has all moments finite. When ε has only a few finite moments, our theoretical convergence rate seems to fall behind the observed rate. Hence, it might be possible to tighten the convergence rate further when ε has only a small number of finite moments.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

M. Anthony and P. L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL https://doi.org/10.1017/CBO9780511624216.

R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:Paper No. 63, 17, 2019. ISSN 1532-4435,1533-7928.

V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. Robust optimization for deep regression, 2015. URL https://arxiv.org/abs/1505.06606.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

G. Grimmett and D. Welsh. *Probability—an introduction*. Oxford University Press, Oxford, second edition, 2014. ISBN 978-0-19-870997-8.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 0-387-95441-4. doi: 10.1007/b97848. URL https://doi.org/10.1007/b97848.

B. Hanin and M. Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2): 251–257, 1991.

D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

Y. Jiao, G. Shen, Y. Lin, and J. Huang. Deep nonparametric regression on approximate manifolds: nonasymptotic error bounds with polynomial prefactors. *Ann. Statist.*, 51(2):691–716, 2023. ISSN 0090-5364,2168-8966. doi: 10.1214/23-aos2266. URL https://doi.org/10.1214/23-aos2266.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv.org*, 2017. ISSN 2331-8422.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2019. ISBN 978-1-119-21468-7. Theory and methods (with R).

L. C. Marsh and D. R. Cormier. *Spline regression models*. Number 137. Sage, 2001.

J. R. Munkres. *Topology*. Prentice Hall, Inc., 2 edition, Jan. 2000. ISBN 0131816292. URL http://www.worldcat.org/isbn/0131816292.

F. Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*, 48(4):1875–1897, 2020. ISSN 0090-5364,2168-8966. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

S. Sharma, S. Sharma, and A. Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6 (12):310–316, 2017.

G. Shen, Y. Jiao, Y. Lin, and J. Huang. Robust nonparametric regression with deep neural networks, 2021.

Z. Shen, H. Yang, and S. Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *J. Math. Pures Appl. (9)*, 157:101–135, 2022. ISSN 0021-7824,1776-3371. doi: 10.1016/j.matpur.2021.07.009. URL https://doi.org/10.1016/j.matpur.2021.07.009.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. doi: 10.1007/978-1-4757-2545-2. URL https://doi.org/10.1007/978-1-4757-2545-2. With applications to statistics.

R. Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596. An introduction with applications in data science, With a foreword by Sara van de Geer.

D. Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.