TECHNISCHE UNIVERSITÄT ILMENAU

Fakultät für Informatik und Automatisierung

FG Systemanalyse, FG Neuroinformatik und Kognitive Robotik

DIPLOMARBEIT

"Quantitative analysis of neural networks as universal function approximators"

zur Erlangung des akademischen Grades

Diplom-Informatiker

vorgelegt von

Stephan Trenn

betreut von

Priv.-Doz. Dr.-Ing. habil. P. Otto

Univ.-Prof. Dr.-Ing. habil. H.-M. Groß

Eingereicht: 11. September 2006, Ilmenau

Inventarisierungsnummer: 2006-10-04/115/IN00/2211

Zusammenfassung

Das mehrschichtige Perzeptron (MLP) ist ein künstliches neuronales Netzwerk, das oft als allgemeiner Funktionsapproximierer genutzt wird. In dieser Diplomarbeit werden die theoretischen Fähigkeiten und Beschränkungen eines MLPs als Funktionsapproximator untersucht. Das wichtigste Resultat ist die explizite Berechnung der notwendigen Größe eines MLPs, um eine gegebene Approximationsordnung zu erreichen, d.h., dass die Taylorpolynome des entsprechenden Grades übereinstimmen. Außerdem wird die Beziehung zwischen der Approximationsordnung und der Approximationsgüte untersucht und es werden Bedingungen für MLPs gegeben, für die eine höhere Approximationsordnung äquivalent zu einer höheren Approximationsgüte ist. Simulationen geben einen ersten Eindruck von der praktischen Relevanz der theoretischen Resultate.

Abstract

The multilayer perceptron (MLP) is an artificial neural network which is widely used as a general function approximator. In this diploma thesis the theoretical capabilities and limits of an MLP as function approximator are studied. The main result is the explicit calculation of the necessary size for an MLP to achieve a given approximation order, i.e. the Taylor polynomials of the corresponding degree coincide. Furthermore, the relation between approximation order and approximation accuracy is studied and conditions for MLPs are given for which a higher approximation order is equivalent to a higher approximation accuracy. Simulations give a first impression of the practical relevance of the theoretical results.

Thesen

- Künstliche neuronale Netze sind Modelle biologischer neuronaler Netze und werden erfolgreich für Anwendungen eingesetzt, in denen Adaptivität und Generalisierungsfähigkeit notwendig ist.
- Das mehrschichtige Perzeptron (MLP) ist ein populäres künstliches neuronales Netzwerk, da es einfach aufgebaut ist und durch den Backpropagation-Algorithmus trainiert werden kann.
- Ein MLP ist in der Lage jede stetige Funktion beliebig genau zu approximieren, wenn es keine Größenbeschränkung für das MLP gibt.
- In praktischen Anwendungen ist die Größe eines MLPs immer beschränkt und es ist wichtig zu wissen, wie die eingeschränkte Größe eines MLPs die theoretisch mögliche Approximationsgüte beeinflusst.
- In einigen Anwendungen ist die Approximationsordnung wichtiger als die globale Approximationsgüte, weil nur die lokale Genauigkeit relevant ist.
- Unter bestimmten Bedingungen sind Approximationsordnung und Approximationsgüte äquivalent.
- **Es ist möglich, die Größe eines MLPs explizit zu bestimmen, welche notwendig ist, um eine gegebene Approximationsordnung zu erreichen.**
- Insbesondere kann entschieden werden, wie viele Schichten ein MLP haben muss.

Theses

- Artificial neural networks are models of biological neural networks and are successfully used for applications where adaptivity and generalization capability is necessary.
- The multilayer perceptron (MLP) is a popular artificial neural network, because it is simple and can be trained with the back-propagation algorithm.
- An MLP is capable of approximating any continuous function arbitrarily well if no restriction on the size of the MLP is made.
- In practical applications the size of an MLP is always bounded and it is important to know how the restricted size of an MLP influences the theoretically possible approximation accuracy.
- In some applications the approximation order is more important than the overall approximation accuracy, because only the local accuracy is relevant.
- Under certain conditions approximation order and approximation accuracy are equivalent.
- **It is possible to explicitly determine the size of an MLP which is necessary to achieve a given approximation order.**
- In particular, it can be decided how many layers an MLP needs to have.

Contents

List of Figures

List of Tables

List of symbols

$\mathbb{N}, \mathbb{Q}, \mathbb{R}$	the natural, rational, and real numbers, resp.
$\mathbb{R}_{\geq 0}, \mathbb{R}_{>0}, \mathbb{N}_{>0}$	the non-negative real numbers, strictly positive real numbers, and strictly positive natural numbers, resp.
$n!$	$= n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$, the factorial for $n \in \mathbb{N}$
$\binom{n}{k}$	$= \frac{n!}{(n-k)!k!}$, the binomial coefficient for $n, k \in \mathbb{N}$

$\mathbf{a} \cdot \mathbf{b}$	$= \sum_{i=1}^{n} a_i b_i$, the standard euclidian inner product for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $n \in \mathbb{N}$, with $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$
h	the number of hidden layers in an MLP, p. 7
n_0	the number of input units in an MLP, p. 7
$\mathbf{n}$	$= (n_0, n_1, \ldots, n_h)$, the number of units per layer in an MLP, p. 7
σ	the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ of an MLP, p. 7
$w_k^{i,j}$	the weight between the j-th unit in layer i and the k-th unit in the layer $i-1$, p. 7
$\theta^{i,j}$	the bias for the j-th unit in the i-th hidden layer, p. 7
$\mathbf{w}^{i,j}$	$= (w_1^{i,j}, w_2^{i,j}, \ldots, w_{n_{i-1}}^{i,j}, \theta^{i,j})$, the weights and the bias belonging to the j-th unit in hidden layer i, p. 7
$\mathbf{W}^i$	$= (\mathbf{w}^{i,1}, \mathbf{w}^{i,2}, \ldots, \mathbf{w}^{i,n_i})$, the weights and biases belonging to hidden layer i, p. 7
$\mathbf{w}^y$	$= (w_1^y, w_2^y, \ldots, w_h^y)$, weights between last hidden layer and the output unit, p. 7
$\mathbf{P}$	$= (\mathbf{W}^1, \mathbf{W}^2, \ldots, \mathbf{W}^h, \mathbf{w}^y)$, all weights and biases, i.e. all parameters, of an MLP, p. 7
$(h, \mathbf{n}, \sigma, \mathbf{P})$	formal definition of an MLP with h hidden layers, $\mathbf{n} = (n_0, n_1, \ldots, n_h)$ units per layers, activation function σ and parameters $\mathbf{P}$, p. 8
$\mathbf{x}$	$= (x_1, x_2, \ldots, x_{n_0})$, the input for an MLP, p. 8
$\mathbf{z}^i$	$= (z_1^i, z_2^i, \ldots, z_{n_i}^i)$, the outputs of the i-th hidden layer, p. 8
y	the output of an MLP, p. 8
f_{MLP}	the MLP function $\mathbf{x} \mapsto y$ for an MLP with fixed parameters $\mathbf{P}$, p. 9
$\text{map}(A \to B)$	the space of all maps from the set A into set B, p. 9

F_{MLP}	the function which maps the parameters to the MLP function, i.e. $\mathbf{P} \mapsto f_{\text{MLP}}$ for some MLP, p. 9
$d(\cdot,\cdot)$	the metric $d : X \times X \to \mathbb{R}_{\geq 0}$ for some metric space X, p. 11 and p. 63
K	some compact set, usually $K = [-1,1]^{n_0}$ for $n_0 \in \mathbb{N}$, p. 12
$C(K \to \mathbb{R})$	the space of all continuous functions $f : K \to \mathbb{R}$, p. 12
$\|f\|$	the maximum norm of $f \in C(K \to \mathbb{R})$, p. 12
$\mathcal{P}_N(K \to \mathbb{R})$	the space of polynomials with degree N or smaller with n_0 variables considered as functions on $K \subseteq \mathbb{R}^{n_0}$, p. 12
$\mathcal{P}(K \to \mathbb{R})$	the space of all polynomials with n_0 variables considered as functions on $K \subseteq \mathbb{R}^{n_0}$, p. 12
$\mathcal{F}_R^{\text{MLP}}(K \to \mathbb{R})$	the space of all MLP functions corresponding to the set of MLPs over a range R, p. 13
$\mathcal{L}(X \to Y)$	the space of continuous linear functions from Banach space X to Banach space Y, p. 17
$\mathcal{L}^n_{\text{sym}}(K \to \mathbb{R})$	the space of symmetric multilinear functions from K^n to $\mathbb{R}$, p. 19
$C^n(K \to \mathbb{R})$	the space of n-times continuously differentiable functions $f : K \to \mathbb{R}$ for $n \in \mathbb{N}$, p. 19
$C^\infty(K \to \mathbb{R})$	the space of arbitrarily often differentiable functions $f : K \to \mathbb{R}$, p. 19
Df or f'	the derivative of $f \in C^1(K \to \mathbb{R})$, p. 17
$D^n f$ or $f^{(n)}$	the n-th derivative of $f \in C^n(K \to \mathbb{R})$ for some $n \in \mathbb{N}$, p. 18
$\mathcal{T}_N\{f\}$	Taylor polynomial at zero of degree $N \in \mathbb{N}$ for $f \in C^N(K \to \mathbb{R})$, p. 21
$C_0^\omega(K \to \mathbb{R})$	the space of functions $f : K \to \mathbb{R}$ which are analytical in zero, p. 24
$\delta_c\{f\}$	radius of convergence of the Taylor series of $f \in C_0^\omega(K \to \mathbb{R})$, p. 24

$C^{\omega}_{\text{nice}}(K \to \mathbb{R})$	the space of nicely analytical functions, p. 25
$\mathcal{CT}_N\{f\}$	the mapping, which maps f to the vector of all coefficients of $\mathcal{T}_N\{f\}$, p. 32
$\mathcal{N}_{\mathcal{CT}}(N, n_0)$	number of coefficients of a polynomial of degree N with n_0 variables, p. 31
$\mathcal{CT}_{N,h,\mathbf{n},\sigma}(\mathbf{P})$	the vector of all coefficients of $\mathcal{T}_N\{f_{\text{MLP}}\}$ for the MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$ considered as function of $\mathbf{P}$, p. 33
$\mathcal{N}_{\text{MLP}}(h, \mathbf{n})$	the number of parameters of an MLP with h hidden layers and $\mathbf{n} = (n_0, n_1, \ldots, n_h)$ units per layer, p. 9
$\mathcal{N}^{h}_{\text{MLP}}(n_0, n)$	the maximal number of parameters for all MLPs with n_0 inputs, n hidden units and h hidden layers (the distribution of the hidden units in the hidden layers is not fixed), p. 38
$\mathcal{N}^{*}_{\text{MLP}}(n_0, n)$	the maximal number of parameters for all MLPs with n_0 inputs and n hidden units (the number of hidden layers is not fixed), p. 38
$\frac{\partial f}{\partial x_i}(\mathbf{x})$	partial derivative of $f \in C^1(K \to \mathbb{R})$ with respect to x_i evaluated at $\mathbf{x} \in K \subseteq \mathbb{R}^{n_0}$, p. 34
$\lceil y \rceil$, $\lfloor y \rfloor$	the floor and ceiling function of $y \in \mathbb{R}$, p. 39

1 Introduction

The original motivation for artificial neural networks (ANNs) was the aim to model cognitive processes observed in animals and humans. Many applications of ANNs show that this approach was very useful, although it also became clear that some problems can not be solved with ANNs or could be solved better with other approaches. There are nowadays lots of different types of ANNs and the connection to biological neural networks is very loose, if there is any at all. For a comprehensive overview over different kinds of neural networks, the interested reader is referred to [Haykin 1994], where also the biological background and historical remarks are given. In this diploma thesis only the multilayer perceptron (MLP) is studied, which is very popular in the application area as well as in theoretical research. The reasons for this popularity might be

- its simplicity,
- its scalability,
- its property to be a general function approximator,
- and its adaptivity.

The MLP was primarily used for classification problems, but its capability to approximate functions made it also interesting for other applications. One of this applications is modelling and control, where artificial neural networks, in particular MLPs, are successfully used (see, e.g., [Nørgaard, Ravn, Poulsen & Hansen 2000]). From an abstract point of view, modelling and (open loop) control with ANNs are very similar. In both cases the ANN should approximate a function, in the first case the function which represents the system which should be modelled and in the second case the function of the inverse system which then can be used as a controller, see [Nørgaard et al. 2000] for more details. When using ANNs in application, there are two main questions:

(i) Is it theoretically possible to solve the task with the considered class of ANNs?

(ii) How can one find an ANN which solves the task?

In general, ANNs are scalable, i.e. they can have different sizes, and they are adaptive, i.e. they have parameters which can be changed. In most cases, the structure and size of an ANN are chosen a priori and afterwards the ANN "learns" a given task, which

is nothing more than adjusting the parameters in a certain way. Therefore, the first question deals with the structure and size of the ANN and the second question targets the change of the parameters, i.e. the learning procedure.

The first question is strongly connected to two other questions:

- What size or structure is necessary to solve a given task?
- What size or structure is sufficient to solve a given task?

This diploma thesis gives an answer to the first of the above two questions for a specific task.

It is an important question whether the necessary size is also sufficient, but an answer to this question is not in the scope of this diploma thesis. The question how to learn an ANN is also not in the scope of this diploma thesis.

The task which is considered here is to approximate any function, which is sufficiently smooth, with a given *approximation order*. One function approximates another function with a specific order if the function value and all derivatives up to the specific order coincide at one fixed point, i.e. the Taylor polynomials are the same (see Figure 1).

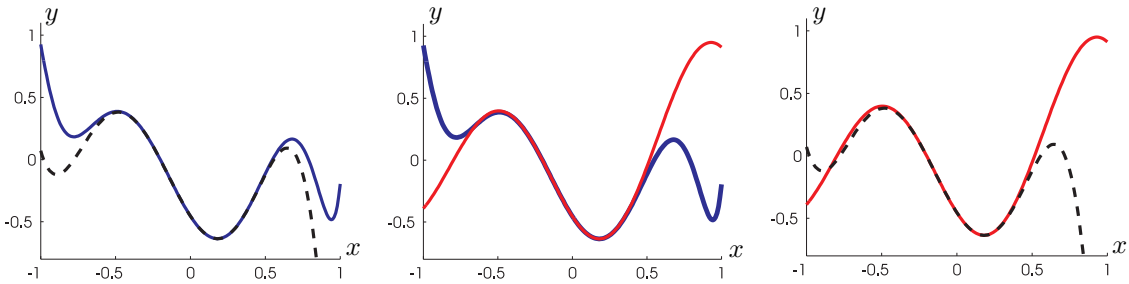

Figure 1: Left and right: Two functions with the same Taylor polynomial of degree five (dashed line), Middle: Both function approximate each other with approximation order five.

This kind of approximation plays an important role in control theory, where often a steady-state is considered and it is important that in a neighbourhood of this steady-state the function which approximates the system or the controller is very accurate. On the other hand, the accuracy far away from the steady-state does not play an important role. The question which will be answered in this diploma thesis is therefore:

Which size and structure is necessary for an MLP to approximate any sufficiently smooth function with a given approximation order?

It is important to highlight here that the size and structure of the MLP should only depend on the given order and not on the specific function which should be learned. This is in contrast to results in [Barron 1994], where the approximation accuracy depends on the specific function which should be approximated. Some results more in the spirit of this diploma thesis are given in [Pinkus 1999], but the bounds of the approximation accuracy could not be used to exactly calculate the necessary size of an MLP, due to unspecified constants in the formulas. The FAQs (frequently asked questions) of the newsgroup `comp.ai.neural-nets` also considers this question and states "In most situations, there is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each." [Sarle 2002].

The diploma thesis is structured as follows. In the second section, the multilayer perceptron (MLP) and its notation is introduced. The third section studies the general properties of the MLP. One important result is that MLPs are capable of approximating any continuous function arbitrarily well, but to achieve this accuracy an arbitrarily big size of the MLP might be necessary. For the approximation order Taylor polynomials play an important role, hence the theory behind Taylor polynomials is elaborated in Section 4. One main result of this section is the correspondence between approximation accuracy and approximation order (Theorem 4.5.1). This result is then used to give conditions for an MLP for which a high approximation order is equivalent to a high approximation accuracy. It turns out that the standard sigmoid activation function of an MLP does not fulfill these conditions, therefore other activation function candidates are proposed.

The main part of this diploma thesis is Section 5, where as the main result Theorem 5.5.2 gives an explicit formula for the size of an MLP which is necessary to achieve a given approximation order for all sufficiently smooth functions. In particular, the number of hidden layers is stated explicitly. The necessary size (i.e. number of units per hidden layer) is given for a range of input dimensions and approximation orders in the appendix. To the author's best knowledge, any explicit results for the necessary size of an ANN to solve a well defined task are not available yet. There seem to be only heuristics and rules of thumb or asymptotic results in the literature.

Finally, some numerical simulations are presented, in the first part with the aim to show how the functions of MLPs are approximated by its own Taylor polynomials for different activation functions and different network sizes. It is clear that if the MLP is not approximated well by its own Taylor polynomial, an approximation of arbitrary functions will in general be worse. In the second part, different MLPs are trained with the standard back-propagation algorithm (first proposed in [Rumelhart, Hinton & Williams 1986]). Firstly, different learning pattern distributions are used to show how this influences the approximation, in particular the approximation order. Secondly, different MLPs (i.e. different size and different activation functions) are trained to give an impression how the different activation functions influence the learning performance. In addition the size of the MLP was chosen much smaller and much bigger than the necessary size, to illustrate the influence of the size on the approximation accuracy.

In the appendix, some mathematical background is collected, as well as some proofs which are not essential for the first reading of the diploma thesis.

2 The multilayer perceptron (MLP)

2.1 Definition of the MLP

The multilayer perceptron (MLP) is a very simple model of biological neural networks and is based on the principle of a feed-forward-flow of information, i.e. the network is structured in an hierarchical way. The MLP consists of different layers where the information flows only from one layer to the next layer. Layers between the input and output layer are called hidden layers, because the units in the hidden layers (the hidden units) are "hidden" from the environment, which only interacts with the input and output units. Note that in the literature, because of the biological background, instead of the term "unit" also the term "neuron" is used. The overall structure of an MLP is illustrated in Figure 2.

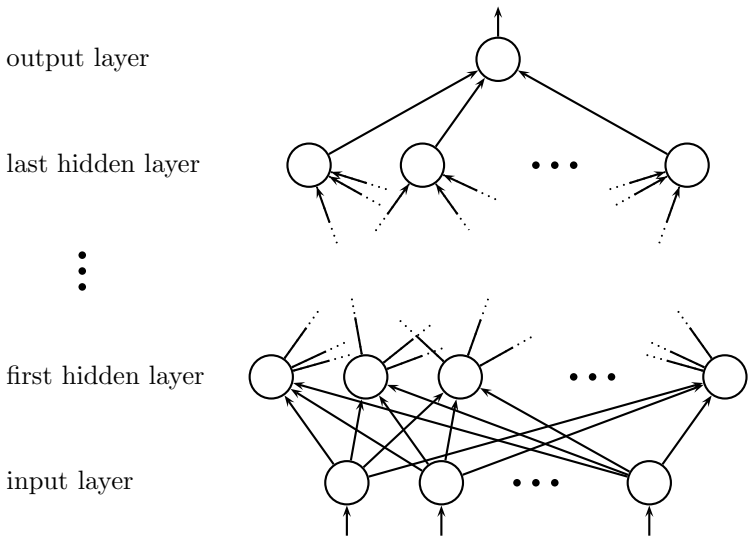

Figure 2: Structure of an MLP. The information flows from the bottom to the top and is processed in each layer by the units in a specific way.

From a theoretical point of view, it is not necessary to consider more than one output unit because two or more output units could be realized by considering two or more MLPs in parallel. However, if the outputs are correlated it may be possible to achieve the same approximation results with fewer hidden units. Nevertheless, a correlation analysis of different outputs and its implications to the necessary number of hidden units is beyond the scope of this work.

The input units play no active role in processing the information flow, because they just distribute the signals to the units of the first hidden layer. All hidden units work in an identical way and the output unit is a simpler version of a hidden unit. In an MLP, each hidden unit transforms the signals from the former layer to one output signal, which is distributed to the next layer. There are two main features of an MLP:

(i) Every edge between two units has a weight.

(ii) Each hidden unit has an, in general nonlinear, activation function.

The activation function is modulo a translation via an individual bias the same for all hidden units. The output of a hidden unit is determined by the weighted sum of the signals from the former layer, which is then transformed by the activation function. The behaviour of a hidden unit is illustrated in the left part of Figure 3. In the output unit the activation function is the identity function, its behaviour is illustrated in the right part of Figure 3. The notation in Figure 3 is explained in the next subsection.

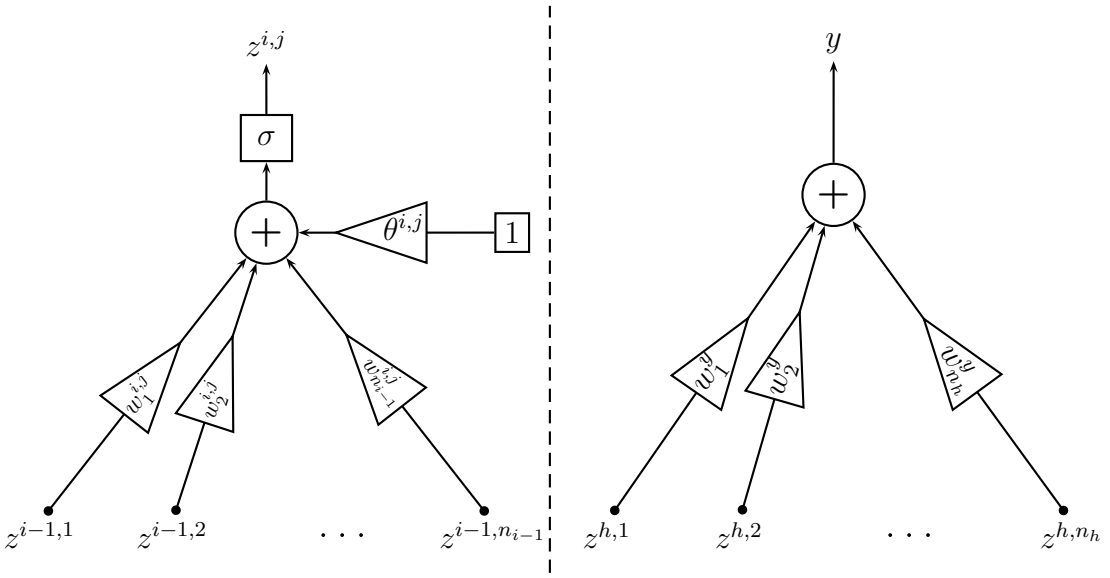

Figure 3: The behaviour of the j-th unit in the i-th hidden layer (left) and of the output unit (right).

2.2 Notation

To make the notation clearer an example is considered first (see Figure 4). For con-

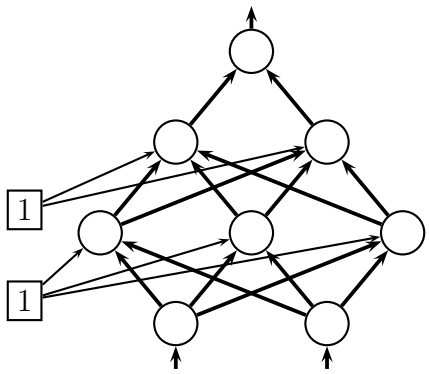

Figure 4: Example of an MLP

venience, the bias is interpreted as an additional unit whose output equals one. The specific value of the biases are then the edges' weights.

The number of hidden layers is denoted by h, for this example $h = 2$. The number of units per layer is $\mathbf{n} = (n_0, n_1, \ldots, n_h)$, where n_0 is the number of input units and n_i, $i \geq 1$, is the number of units in the i-th hidden layer. Here, $\mathbf{n} = (2, 3, 2)$. The activation function is $\sigma : \mathbb{R} \to \mathbb{R}$, and, for the standard MLP, is given by

$$\sigma(t) = \frac{1}{1 + e^{-t}}.$$

For the results in this diploma thesis, the specific form of the activation is not relevant and most results hold true for any activation function with some qualitative properties.

The above parameters of an MLP are in general chosen a priori and are in most cases not changed while the network is "learning". In the learning process of an MLP the variable parameters are adapted in a specific way. These variable parameters, which consist of the edges' weights and the biases, are summarized in $\mathbf{P}$. In detail,

$$\mathbf{P} = (\mathbf{W}^1, \mathbf{W}^2, \ldots, \mathbf{W}^h, \mathbf{w}^y),$$

where $\mathbf{w}^y = (w_1^y, \ldots, w_{n_h}^y)$ are the n_h weights for the output unit and all weights and biases for the i-th hidden layer, $1 \leq i \leq h$, are collected in $\mathbf{W}^i = (\mathbf{w}^{i,1}, \mathbf{w}^{i,2}, \ldots, \mathbf{w}^{i,n_i})$. The component $\mathbf{w}^{i,j} = (w_1^{i,j}, w_2^{i,j}, \ldots, w_{n_{i-1}}^{i,j}, \theta^{i,j})$, for $1 \leq i \leq h$ and $1 \leq j \leq n_i$, represents all weights and the bias which belong to the j-th unit in the i-th hidden layer. For the given example these parameters are

$$\mathbf{P} = (\mathbf{W}^1, \mathbf{W}^2, \mathbf{w}^y),$$

$$
\begin{aligned}
\mathbf{W}^1 &= (\mathbf{w}^{1,1}, \mathbf{w}^{1,2}, \mathbf{w}^{1,3}),\\
&\quad \mathbf{w}^{1,1} = (w_1^{1,1}, w_2^{1,1}, \theta^{1,1}) \in \mathbb{R}^3,\\
&\quad \mathbf{w}^{1,2} = (w_1^{1,2}, w_2^{1,2}, \theta^{1,2}) \in \mathbb{R}^3,\\
&\quad \mathbf{w}^{1,3} = (w_1^{1,3}, w_2^{1,3}, \theta^{1,3}) \in \mathbb{R}^3,\\
\mathbf{W}^2 &= (\mathbf{w}^{2,1}, \mathbf{w}^{2,2}),\\
&\quad \mathbf{w}^{2,1} = (w_1^{2,1}, w_2^{2,1}, w_3^{2,1}, \theta^{2,1}) \in \mathbb{R}^4,\\
&\quad \mathbf{w}^{2,2} = (w_1^{2,2}, w_2^{2,2}, w_3^{2,2}, \theta^{2,2}) \in \mathbb{R}^4,\\
\mathbf{w}^y &= (w_1^y, w_2^y) \in \mathbb{R}^2.
\end{aligned}
$$

Denote the output of the j-th unit in the i-th hidden layer by $z^{i,j}$ and collect all outputs of the layer i in the vector $\mathbf{z}^i = (z^{i,1}, \ldots, z^{i,n_i}, 1)$, where the last component stands for the virtual bias unit. The n_0-dimensional input of the MLP is $\mathbf{x} = (x_1, x_2, \ldots, x_{n_0})$ and the output is y. For notational convenience consider the input layer as the zeroth hidden layer, i.e. $\mathbf{z}^0 = (x_1, \ldots, x_{n_0}, 1)$.

Finally the formal definition of an MLP is given:

Definition 2.2.1 (Multilayer Perceptron - MLP)**.** *A* multilayer perceptron *(MLP) is a quadtuple*

$$(h, \mathbf{n}, \sigma, \mathbf{P}),$$

where $h \in \mathbb{N}$ is the number of hidden layers, $\mathbf{n} = (n_0, n_1, \ldots, n_h) \in \mathbb{N}^{h+1}$ is the number of units per hidden layer (the hidden layer zero is the input layer), $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function and

$$\mathbf{P} = (\mathbf{W}^1, \ldots, \mathbf{W}^h, \mathbf{w}^y),$$

where, for $i = 1, \ldots, h$, $\mathbf{W}^i = (\mathbf{w}^{i,1}, \ldots, \mathbf{w}^{i,n_i}) \in (\mathbb{R}^{n_{i-1}+1})^{n_i}$ are the parameters (weights and biases) between the $(i-1)$-th and i-th hidden layer and $\mathbf{w}^y \in \mathbb{R}^{n_h}$ are the parameters between the last hidden layer and the output unit.

2.3 The MLP function

So far, the formal relation between the output y and the inputs $x_1, x_2, \ldots, x_{n_0}$ was not specified. Consider therefore a fixed MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$, in particular the "variable parameters" $\mathbf{P}$ are fixed. The main idea is to view the MLP as a black box with n_0

inputs and one output. In the former subsection the signal flow and transformation was already informally described, the formal definition is as follows.

Definition 2.3.1 (MLP function)**.** *For an MLP* $(h, \mathbf{n}, \sigma, \mathbf{P})$ *as in Definition 2.2.1, the* MLP-function

$$f_{MLP} : \mathbb{R}^{n_0} \to \mathbb{R}, \quad \mathbf{x} = (x_1, x_2, \ldots, x_{n_0}) \mapsto y$$

is recursively defined by

$$y = \mathbf{w}^y \cdot \mathbf{z}^h, \text{ where}$$

$$\begin{aligned}
\mathbf{z}^h &= \big(\sigma(\mathbf{w}^{h,1} \cdot \mathbf{z}^{h-1}) \ , \ \sigma(\mathbf{w}^{h,2} \cdot \mathbf{z}^{h-1}) \ , \ \ldots \ , \ \sigma(\mathbf{w}^{h,n_h} \cdot \mathbf{z}^{h-1}) \big), \\
\mathbf{z}^i &= \big(\sigma(\mathbf{w}^{i,1} \cdot \mathbf{z}^{i-1}) \ , \ \ldots \ , \ \sigma(\mathbf{w}^{i,n_i} \cdot \mathbf{z}^{i-1}) \ , \ 1 \big) \quad \text{for } i = h-1, \ldots, 1, \\
\mathbf{z}^0 &= (x_1, x_2, \ldots, x_{n_0}, 1).
\end{aligned}$$

Note that two MLPs with the same structure but with different parameters $\mathbf{P}_1$ and $\mathbf{P}_2$ have in general different MLP functions. Indeed, one could consider another more abstract mapping, which maps each parameter set $\mathbf{P}$ to the corresponding MLP function. Let $\mathcal{N}_{\text{MLP}}(h, \mathbf{n})$ be the number of parameters in $\mathbf{P}$, where h and $\mathbf{n}$ determine the fixed structure of the MLP, and let $\text{map}(A \to B)$ be the space of mappings from a set A to a set B. This abstract mapping is then denoted by

$$F_{\text{MLP}} : \mathbb{R}^{\mathcal{N}_{\text{MLP}}(h,\mathbf{n})} \to \text{map}(\mathbb{R}^{n_0} \to \mathbb{R}), \quad \mathbf{P} \mapsto F_{\text{MLP}}[\mathbf{P}] = f_{\text{MLP}},$$

where f_{MLP} is the MLP function of the MLP with the specific parameter set $\mathbf{P}$. Square brackets are used because $F_{\text{MLP}}[\mathbf{P}]$ is itself a function. In order to highlight the dependence between the MLP function f_{MLP} and the MLP parameters $\mathbf{P}$ it is now possible to write $F_{\text{MLP}}[\mathbf{P}](\mathbf{x})$ instead of $f_{\text{MLP}}(\mathbf{x})$.

3 Qualitative properties of the MLP as general function approximator

This section is mainly based on the survey paper [Pinkus 1999].

3.1 Mathematical preliminaries

From an abstract point of view, approximation deals with a set of objects, say O, which might be quite complicated, and a subset $A \subset O$, which is easier to handle. The aim is to find for each complicated object $o \in O$ a simpler object $a \in A$, which is in some sense close to o. As an example consider as the complicated set $O = \mathbb{R}$, the real numbers, and as the simpler set $A = \mathbb{Q}$, the rational numbers. Clearly it is possible to find for each real number $x \in \mathbb{R}$ a rational number $q \in \mathbb{Q}$ such that $|x - q|$ is small.

To make the term "close to" precise in the general setting, the distance between two arbitrary objects must be defined. In a mathematical context, the set O must be assumed to be a metric space (O, d), where $d : O \times O \to \mathbb{R}_{\geq 0}$ is a metric and $d(o_1, o_2)$ is the distance between two objects $o_1, o_2 \in O$. See the appendix for details on metric spaces. For real numbers, the standard metric is defined as $d(x, y) = |x - y|$ for any $x, y \in \mathbb{R}$, but it is possible to define other metrics, e.g., $d(x, y) = |x - y|/(1 + |x - y|)$ which only has values in the interval $[0, 1)$.

The following formal definition of denseness plays an important role in approximation theory.

Definition 3.1.1 (Denseness)**.** *Let (O, d) be a metric space and $A \subseteq O$.*
For $\varepsilon > 0$ the set A is called ε-dense *in O if, and only if,*

$$\forall\, o \in O\ \exists\, a \in A : \quad d(o, a) < \varepsilon.$$

The set A is called dense *in O if, and only if, A is ε-dense in O for every $\varepsilon > 0$.*

The concept of ε-denseness is particularly relevant for numerical approximations with computers. Because of the finite machine precision, an arbitrary good approximation is in most cases not possible or needed. The set of rational numbers as a subset of the real numbers is a typical example for a dense set.

An important metric space for this diploma thesis is the set of continuous functions, denoted by

$$C(K \to \mathbb{R}) := \{\ f : K \to \mathbb{R}\ \mid f \text{ is continuous }\},$$

where $K \subseteq \mathbb{R}^{n_0}$ is some compact set (i.e. bounded and closed, see appendix), together with the metric

$$d : C(K \to \mathbb{R}) \times C(K \to \mathbb{R}) \to \mathbb{R}, \quad (f,g) \mapsto d(f,g) := \max_{\mathbf{x} \in K} \big|f(\mathbf{x}) - g(\mathbf{x})\big|.$$

Note that this metric space is also a normed space, see the appendix for more details. The distance between two functions is therefore also denoted by $\|f-g\|$ which is equal to $d(f,g)$.
An important subspace of the continuous functions $C(K \to \mathbb{R})$ is the space of polynomials of degree $N \in \mathbb{N}$, considered as functions on $K \subseteq \mathbb{R}^{n_0}$,

$$\mathcal{P}_N(K \to \mathbb{R}) := \left\{ p : K \to \mathbb{R} \;\middle|\; \begin{array}{l} p(\mathbf{x}) = \displaystyle\sum_{|I| \le N} a_I \mathbf{x}^I, \\ \quad a_I \in \mathbb{R} \text{ for all } I \in \mathbb{N}^{n_0} \text{ with } |I| \le N \\ \quad \text{and } \mathbf{x} = (x_1, x_2, \ldots, x_{n_0}) \in K \end{array} \right\},$$

where $\mathbf{x}^I = {x_1}^{i_1} {x_2}^{i_2} \cdots {x_{n_0}}^{i_{n_0}}$ for $I = (i_1, i_2, \ldots, i_{n_0}) \in \mathbb{N}^{n_0}$ and $|I| = i_1 + i_2 + \ldots + i_{n_0}$.
The space of all polynomials is

$$\mathcal{P}(K \to \mathbb{R}) := \bigcup_{N \in \mathbb{N}} \mathcal{P}_N(K \to \mathbb{R}).$$

The subsection is finished with interesting density properties of polynomials in relation to continuous functions.

Proposition 3.1.2 (Denseness of polynomials)**.**

(i) $\mathcal{P}(K \to \mathbb{R})$ *is dense in* $C(K \to \mathbb{R})$*, i.e. for every continuous function* f *and every arbitrarily small* $\varepsilon > 0$ *there exists a polynomial* p *such that* $\|f - p\| < \varepsilon$*.*

(ii) For a fixed degree $N \in \mathbb{N}$*, the space* $\mathcal{P}_N(K \to \mathbb{R})$ *is* not *ε-dense in* $C(K \to \mathbb{R})$ *for any arbitrarily large* $\varepsilon > 0$*.*

The proposition is proved in the appendix.

3.2 The general approximation capability of MLPs

MLPs are used to approximate given functions. Even if MLPs are used for classification problems, this can be viewed as an approximation of a function, namely the identification function. Depending on the considered problem, there is a set $\mathcal{F}$ of possible functions, which should be approximated. For example, $\mathcal{F} = C(\mathbb{R}^{n_0} \to \mathbb{R})$ for continuous approximation problems or $\mathcal{F} = \operatorname{map}\big(\mathbb{R}^{n_0} \to \{1, 2, \ldots, N\}\big)$ for classification problems. The latter can be viewed as an special case of the former by identifying whole intervals as classes, i.e. each input whose function value lies in a certain interval belongs to the same class. Functions arising from technical systems can be assumed to be continuous in most cases. In addition, there often exist physical or other bounds for the input, hence it is reasonable to consider $\mathcal{F} = C(K \to \mathbb{R})$, where $K \subseteq \mathbb{R}^{n_0}$ is a compact subset of the n_0-dimensional input space. It is no restriction to assume that $K = [-1, 1]^{n_0}$, i.e. the input signals are scaled such that $x_1, x_2, \ldots, x_{n_0} \in [-1, 1]$.

The approximation capability of an MLP is governed by the space of possible MLP functions. It is not very fruitful to allow for all possible MLPs, instead it is necessary to restrict oneself to MLPs with some structural assumption. The following notation is used:

$$\mathcal{F}^{\mathrm{MLP}}_{(\alpha_h, \alpha_{\mathbf{n}}, \alpha_\sigma, \alpha_{\mathbf{P}})}(K \to \mathbb{R}) := \left\{ f : K \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ is the MLP function of an MLP } (h, \mathbf{n}, \sigma, \mathbf{P}) \\ \text{with } h \text{ arbitrary if } \alpha_h = \cdot \text{ or } h = \alpha_h \text{ otherwise,} \\ \text{and likewise for } \mathbf{n}, \sigma \text{ and } \mathbf{P} \end{array} \right\}.$$

To make this notation clearer, consider the following examples:

- $\mathcal{F}^{\mathrm{MLP}}_{(\cdot,\cdot,\cdot,\cdot)}(K \to \mathbb{R})$ is the space of all possible MLP functions with arbitrary activation functions, number of hidden layers and units, weights and biases.
- $\mathcal{F}^{\mathrm{MLP}}_{(1,\cdot,\sigma,\cdot)}(K \to \mathbb{R})$ is the space of all MLP functions of MLPs with one hidden layer and the activation function σ.
- $\mathcal{F}^{\mathrm{MLP}}_{(2,(n_0,n_1,n_2),\sigma,\cdot)}(K \to \mathbb{R})$ is the space of all MLP functions of MLPs with the activation function σ and a fixed structure with two hidden layers and n_1 units in the first hidden layer and n_2 units in the second hidden layer.

- $\mathcal{F}^{\text{MLP}}_{(h,\mathbf{n},\sigma,\mathbf{P})}(K \to \mathbb{R})$ consists only of the single MLP function of the MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$.

It is now possible to formulate a first positive result:

Theorem 3.2.1 (Denseness of single hidden layer MLPs)**.** *Let σ be continuous and not a polynomial, then $\mathcal{F}^{MLP}_{(1,\cdot,\sigma,\cdot)}(K \to \mathbb{R})$ is dense in $C(K \to \mathbb{R})$, i.e. for every continuous function $f : K \to \mathbb{R}$ and every arbitrarily small $\varepsilon > 0$ there exists an MLP $(1, \mathbf{n}, \sigma, \mathbf{P})$ with one hidden layer such that its MLP function f_{MLP} fulfills $\|f - f_{MLP}\| < \varepsilon$.*

The theorem is proved in the appendix.

One might wonder about the assumption that σ must not be a polynomial. The reason is that if σ was a polynomial of some finite degree, the MLP function would be a polynomial of the same degree and hence, by Proposition 3.1.2, denseness is not possible.

Although this result looks promising it has a huge drawback. The MLP structure, namely the number of hidden units, depends on the function, which should be approximated, and on the accuracy ε. It is therefore not possible to conclude from Theorem 3.2.1 that there exists an MLP with a fixed structure which is able to approximate all continuous functions with an arbitrary accuracy. The next subsection considers this question in detail.

3.3 MLPs with a fixed network structure

For implementing neural networks it is important that the structure of an MLP is fixed (e.g., as an integrated circuit) and only the weight and bias parameters are adapted. For single hidden layer MLPs there seems to be a negative result.

Conjecture 3.3.1 (Non-ε-denseness of MLPs)**.** *$\mathcal{F}^{MLP}_{(1,\mathbf{n},\sigma,\cdot)}(K \to \mathbb{R})$ is not ε-dense in $C(K \to \mathbb{R})$ for every $\varepsilon > 0$ and every fixed number $\mathbf{n}$ of hidden units, i.e. for a given accuracy it is not possible to find an MLP with a fixed structure which can approximate all continuous function with the given accuracy.*

The conjecture is backed up by [Pinkus 2006], but it still seems to be an open problem.
In view of the above conjecture one might ask if things get better if one considers more than one hidden layer. Indeed, there is a positive result for two hidden layers.

Proposition 3.3.2 (Two hidden layers)**.** *There exists an activation function σ such that for $\mathbf{n} = (n_0, 2n_0 + 1, 4n_0 + 3)$ the set $\mathcal{F}^{MLP}_{(2,\mathbf{n},\sigma,\cdot)}(K \to \mathbb{R})$ is dense in $C(K \to \mathbb{R})$, i.e. there exists a two hidden layers MLP with a fixed structure that can approximate any continuous function arbitrarily well.*

The proposition is proved in the appendix.

Unfortunately, the needed activation function is very complex and can not be implemented in any real MLP and therefore the result is only of theoretical interest.

3.4 Conclusion for approximation with MLPs

For the function space $\mathcal{F} = C(K \to \mathbb{R})$ it was shown that in general MLPs are capable of approximating functions in $\mathcal{F}$ arbitrarily well. But if one restricts the MLPs to a fixed structure, which is inevitable for applications, there seems to be no possibility to achieve good approximation results for $\mathcal{F}$. The problem is that the space of continuous functions is still too large, a general continuous function can exhibit really strange behaviour. It can be shown that if differentiability and boundedness of the derivatives are assumed, then it is possible to achieve ε-boundedness with a fixed MLP structure. Because the results are very technical and do not allow for a calculation of the necessary hidden units they are only considered in the appendix.
In the next section, instead of continuous functions, the "nicer" analytical functions are considered with the aim to obtain quantitative approximation results.

4 Taylor polynomials and function approximation

4.1 Mathematical preliminaries

In the following, derivatives of functions play an important role, therefore a brief summary of derivatives in general Banach spaces is given (Banach spaces are complete normed vector spaces, see the appendix for details). Let X and Y be Banach spaces and $f : X \to Y$. The derivative $Df(\mathbf{x})$ of f at some point $\mathbf{x} \in X$ is a continuous linear function from X to Y, with

$$\lim_{h \to 0} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})(\mathbf{h})}{\|\mathbf{h}\|} = 0.$$

Note that $Df(\mathbf{x})$ is itself a function, therefore the notation $Df(\mathbf{x})(\mathbf{h})$ makes sense. Introducing the notation $\mathcal{L}(X \to Y)$ for the space of all continuous linear functions from X to Y, the derivative of f is

$$Df : X \to \mathcal{L}(X \to Y), \quad \mathbf{x} \mapsto Df(\mathbf{x}),$$

if f is differentiable. Since this viewpoint of derivatives is not very common consider the following example

$$f : \mathbb{R}^2 \to \mathbb{R}^3, \quad (x_1, x_2) \mapsto \begin{pmatrix} x_1 + x_2 \\ x_1^2 \\ \sin x_2 \end{pmatrix}.$$

The derivative at the point $\mathbf{x} = 0$ is

$$Df(0) = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 2},$$

and can be viewed as an linear mapping in form of a matrix-vector-multiplication:

$$Df(0) : \mathbb{R}^2 \to \mathbb{R}^3, \quad (h_1, h_2) \mapsto Df(0) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} h_1 + h_2 \\ 0 \\ h_2 \end{pmatrix}.$$

The derivative $Df(\mathbf{x})$ at some arbitrary point $\mathbf{x} \in \mathbb{R}^2$ is

$$Df(\mathbf{x}) = \begin{bmatrix} 1 & 1 \\ 2x_1 & 0 \\ 0 & \cos x_2 \end{bmatrix} \in \mathbb{R}^{3\times 2},$$

which is for every fixed $\mathbf{x} \in \mathbb{R}^2$ a 3×2 matrix, which can be viewed as a linear mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$. Note that Df as a function of $\mathbf{x} \in X$ is in general not linear.
On a first glance the above definition for derivatives seems to collide with the "normal" derivative of one-dimensional functions $f : \mathbb{R} \to \mathbb{R}$, because there $Df : \mathbb{R} \to \mathbb{R}$ instead of $Df : \mathbb{R} \to \mathcal{L}(\mathbb{R} \to \mathbb{R})$. But it is easy to see that $\mathbb{R}$ and $\mathcal{L}(\mathbb{R} \to \mathbb{R})$ can be identified with each other by interpreting a real number $a \in \mathbb{R}$ as a 1×1-matrix which, as above, can be viewed as a linear mapping from $\mathbb{R}$ to $\mathbb{R}$.

Things get a bit more complicated if one considers higher derivatives. First observe that the space $Y_1 := \mathcal{L}(X \to Y)$ is a Banach space again with the norm

$$\|L\| := \sup_{\mathbf{x}\in X\setminus\{0\}} \frac{\big\|L(\mathbf{x})\big\|}{\|\mathbf{x}\|} \tag{4.1.1}$$

for any $L \in \mathcal{L}(X \to Y)$. The above definition of the derivative can now be used to obtain the second derivative by taking the derivative of $Df : X \to Y_1$:

$$D(Df) : X \quad \to \quad \mathcal{L}(X \to Y_1) = \mathcal{L}\big(X \to \mathcal{L}(X \to Y)\big).$$

Note that $\mathcal{L}\big(X \to \mathcal{L}(X \to Y)\big)$ can be interpreted as a subset of $\operatorname{map}(X^2 \to Y)$. Clearly this process can now be repeated to obtain higher derivatives. Formally define

$$\begin{aligned} &D^0 f := f \quad \text{and} \quad D^n f := D(D^{n-1} f), \\ &Y_0 := Y \quad \text{and} \quad Y_n := \mathcal{L}(X \to Y_{n-1}). \end{aligned} \tag{4.1.2}$$

Hence

$$D^n f : X \to Y_n,$$

and in particular $D^n f(\mathbf{x}) \in Y_n$ for every $\mathbf{x} \in X$, where Y_n can be interpreted as a subset of $\operatorname{map}(X^n \to Y)$. Instead of $D^n f$ the notation $f^{(n)}$ or, if $n \leq 3$, f', f'', and f''', is also used.

In this diploma thesis the focus is on functions $f : K \to \mathbb{R}$, i.e. $X = K \subseteq \mathbb{R}^{n_0}$ and $Y = \mathbb{R}$, in this case elements in Y_1 can be interpreted as row vectors (the gradient) and elements in Y_2 as matrixes (the Hessian matrix). For higher derivatives the elements in Y_n can be viewed as tensors of rank n.
For $\mathbf{x} \in K$ and $D^n f(\mathbf{x}) \in Y_n$, where Y_n is interpreted as a subset of $\mathrm{map}(X^n \to Y)$, write

$$D^n f(\mathbf{x})\mathbf{h}^n := D^n f(\mathbf{x})(\mathbf{h}, \mathbf{h}, \ldots, \mathbf{h}).$$

Note that for a fixed $\mathbf{x} \in K$

$$D^n f(\mathbf{x})\mathbf{h}^n = \sum_{i_1, i_2, \ldots, i_n = 1}^{n} \frac{\partial^n f}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_n}}(\mathbf{x}) h_{i_1} h_{i_2} \cdots h_{i_n},$$

i.e. $D^n f(\mathbf{x})\mathbf{h}^n$ is for a fixed $\mathbf{x} \in \mathbb{K}$ a multivariable polynomial with variables $\mathbf{h} = (h_1, h_2, \ldots, h_n)$, which monomials all have degree n.

If $D^n f(\mathbf{x})$ exists for all $\mathbf{x} \in K$ and Df is a continuous function (in $\mathbf{x}$), then f is called n-times continuously differentiable on K. The space of all n-times continuously differentiable functions on K is denoted by $C^n(K \to \mathbb{R})$ and the space $C^\infty(K \to \mathbb{R})$ is the space of arbitrarily often differentiable functions, which are also called smooth functions.

It is well known that for partial derivatives of continuously differentiable function it is not important in which order the partial derivatives are calculated. Formally this can be expressed as (see [Amann & Escher 2001*b*, Kor. VII.5.3])

$$D^n f(\mathbf{x}) \in \mathcal{L}^n_{\mathrm{sym}}(K \to \mathbb{R}),$$

where $\mathcal{L}^n_{\mathrm{sym}}(X \to Y)$ is a subspace of the space Y_n as defined in (4.1.2) with the additional condition that for all $L \in \mathcal{L}^n_{\mathrm{sym}}(X \to Y)$, for all $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in X$, and for all permutations $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$

$$L(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = L(\mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)} \ \ldots, \mathbf{x}_{\pi(n)}).$$

Finally, multivariable power series are considered:

Definition 4.1.1 (Multivariable power series)**.** *Let X be a Banach space. A multi-*

variable power series P is defined by

$$P(\mathbf{x}) = \sum_{k=0}^{\infty} A_k(\underbrace{\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x}}_{k \text{ times}}),$$

where $\mathbf{x} \in X$ and $A_k \in \mathcal{L}^k_{sym}(X \to \mathbb{R})$, $k \in \mathbb{N}$, are called coefficients of P

Multivariable power series are needed to study Taylor series in the next section, where the following important property is used:

Proposition 4.1.2. *Let P be a multivariable power series on $K = [-1, 1]^{n_0}$ with*

$$\limsup_{k \to \infty} \sqrt[k]{\|A_k\|} < 1$$

for the coefficients $A_k \in \mathcal{L}^k_{sym}(X \to \mathbb{R})$, $k \in \mathbb{N}$. Then $P(\mathbf{x})$ converges absolutely on K and P is arbitrarily often differentiable at $\mathbf{x} = 0$ with

$$D^n P(0) = n! A_n,$$

for all $n \in \mathbb{N}$.

The proof is technical and is therefore put in the appendix.

4.2 Taylor polynomials

A geometric interpretation of derivatives is that the derivative at some point is locally a linear best approximation of the function (see Figure 5), i.e.

$$f(\mathbf{x}) \approx f(0) + Df(0)(\mathbf{x}).$$

The right hand side is an affine linear function or in other terms a polynomial of degree one.

An obvious generalization of the geometric approach is to consider also quadratic, cubic and higher approximations, which yield polynomials of degree two, three or higher. These polynomials are called Taylor polynomials and the following theorem holds.

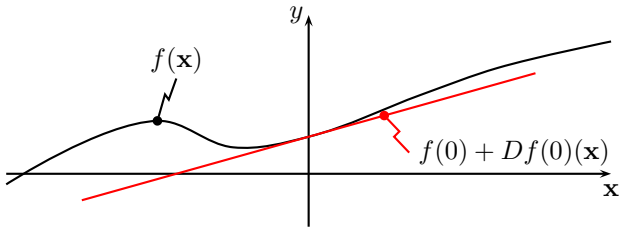

Figure 5: Linear best approximation

Theorem 4.2.1 (Taylor). *Let $N \in \mathbb{N}$, $K = [-1,1]^{n_0}$ and $f \in C^{N+1}(K \to \mathbb{R})$ then*

$$f(\mathbf{x}) = \underbrace{\sum_{k=0}^{N} \frac{D^k f(0)}{k!} \mathbf{x}^k}_{\mathcal{T}_N\{f\}(\mathbf{x})} + R_{N+1}\{f\}(\mathbf{x})$$

where the remainder term $R_{N+1}\{f\}(\mathbf{x})$ fulfills

$$|R_{N+1}\{f\}(\mathbf{x})| \leq \frac{1}{(N+1)!} \max_{\xi \in K} |D^{N+1} f(\xi) \mathbf{x}^{N+1}|.$$

This is a standard result in analysis and a proof can for example be found in [Amann & Escher 2001*b*, Thm. VII.5.8]. Note that $\mathbf{x} \mapsto \mathcal{T}_N\{f\}(\mathbf{x}) \in \mathcal{P}_N(K \to \mathbb{R})$, i.e. the first term is a polynomial of degree N. Under certain conditions, which will be studied later, the remainder term is negligible and $f(\mathbf{x}) \approx \mathcal{T}_N\{f\}(\mathbf{x})$ for all small $\mathbf{x} \in K$. Clearly, $\mathcal{T}_N\{f\}$ can also be defined if f is only N times continuously differentiable, but than the remainder term can not be bounded as in Theorem 4.2.1. To illustrate the theorem, consider the example $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = \sin(x)$. In Figure 6 different Taylor polynomials are plotted.

The Taylor polynomials for this example are given by

$$\begin{aligned}
\mathcal{T}_3\{f\}(x) &= x - \frac{1}{3}x^3 \\
\mathcal{T}_9\{f\}(x) &= x - \frac{1}{3}x^3 + \frac{1}{120}x^5 - \frac{1}{5040}x^7 + \frac{1}{362880}x^9 \\
\mathcal{T}_{15}\{f\}(x) &= \mathcal{T}_9\{f\}(x) - \frac{1}{39916800}x^{11} + \frac{1}{6227020800}x^{13} - \frac{1}{1307674368000}x^{15}
\end{aligned}$$

Note that in this case $n_0 = 1$ and therefore tensors do not appear in the Taylor

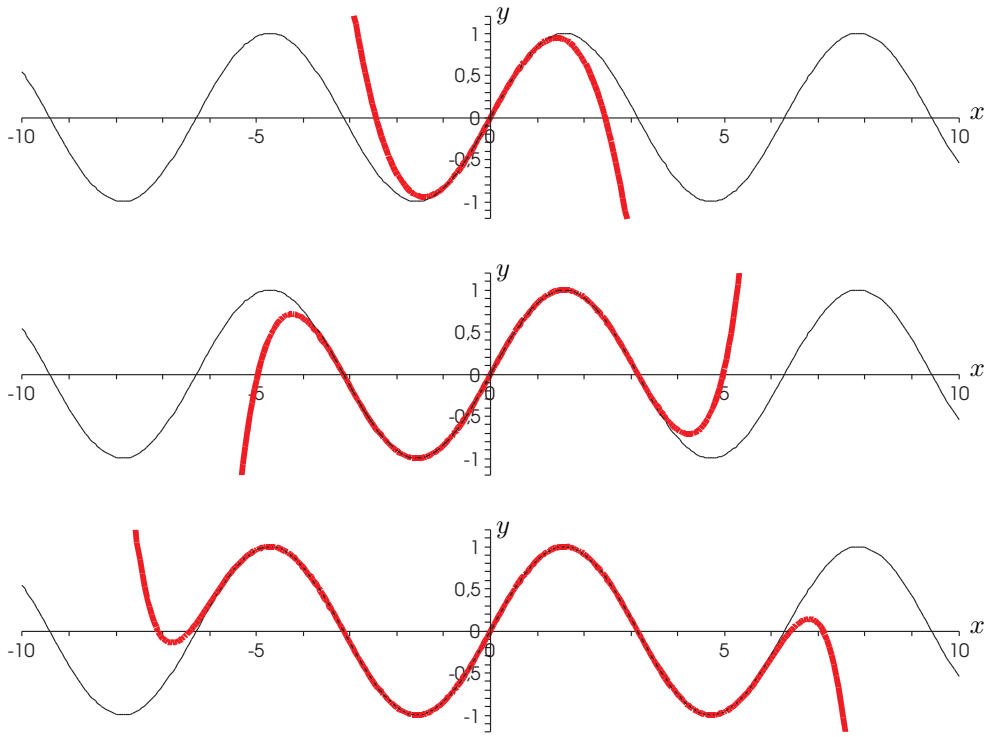

Figure 6: Taylor polynomials (thick lines) for $f(x) = \sin(x)$ of degree 3, 9, 15, resp.

polynomial. To show the effect of tensors consider the example

$$f : \mathbb{R}^3 \to \mathbb{R}, \quad \mathbf{x} = (x_1, x_2, x_3) \mapsto e^{x_1}({x_2}^2 + {x_3}^3).$$

Writing tensors of rank three as a vector of matrixes and a tensor of rank four as matrix with matrix-entries, the first four derivatives of f at $\mathbf{x} = (0, 0, 0)$ can be written as

$$Df(0) = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix},$$

$$D^2 f(0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$D^3 f(0) = \left\{ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6 \end{bmatrix} \right\}$$

and

$$D^4 f(0) = \left\{ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6 \end{bmatrix} \right.$$
$$\begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$\left. \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 6 \\ 0 & 0 & 0 \\ 6 & 0 & 0 \end{bmatrix} \right\}.$$

The Taylor polynomial of degree four is for this example

$$\mathcal{T}_4\{f\}(\mathbf{x}) = {x_2}^2 + x_1{x_2}^2 + {x_3}^3 + \frac{1}{2}{x_1}^2{x_2}^2 + x_1{x_3}^3.$$

4.3 Analytical functions

So far, the accuracy of $\mathcal{T}_N\{f\}$ was not studied. It is a classical result in analysis (e.g., [Amann & Escher 2001*b*]) that the remainder term in Theorem 4.2.1 fulfills

$$\lim_{\mathbf{x}\to 0} \frac{R_{N+1}\{f\}(0)(\mathbf{x})}{\|\mathbf{x}\|^N} = 0, \tag{4.3.1}$$

i.e. $R_{N+1}\{f\}(0)(\mathbf{x}) = o\big(\|\mathbf{x}\|^N\big)$, where $\|\mathbf{x}\|$ is any norm of $\mathbf{x} \in \mathbb{R}^{n_0}$ (see the appendix for details on norms in $\mathbb{R}^n$). In terms of approximation accuracy this implies that there exists for every $N \in \mathbb{N}$ and every $\varepsilon > 0$ a $\delta > 0$, which depends on N and ε, such that

$$\big|f(\mathbf{x}) - \mathcal{T}_N\{f\}(\mathbf{x})\big| < \varepsilon \quad \text{for all } \mathbf{x} \text{ with } \|\mathbf{x}\| < \delta.$$

Unfortunately this does not guarantee good approximation of f in the sense of $\big\|f - \mathcal{T}_N\{f\}\big\| < \varepsilon$ for a given small ε, because therefore a global (i.e. on the whole of K) accuracy is necessary and not only on a small environment of the origin. Clearly global accuracy can only be improved by increasing the degree of the Taylor polynomial, and a necessary condition for arbitrary good global accuracy is

$$\lim_{N\to\infty} \mathcal{T}_N\{f\}(\mathbf{x}) = f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in K.$$

This condition is closely related to analyticity:

Definition 4.3.1 (Analyticity)**.** *A smooth function $f \in C^\infty(K \to \mathbb{R})$ is called analytical (in zero) if, and only if, there exists a $\delta > 0$ such that*

$$\lim_{N\to\infty} \mathcal{T}_N\{f\}(\mathbf{x}) = f(\mathbf{x}) \quad \textit{for all } \mathbf{x} \in (-\delta, \delta)^{n_0}. \tag{4.3.2}$$

The largest δ such that equation (4.3.2) holds is called radius of convergence and is denoted by $\delta_c\{f\}$. The space of all analytical functions from K to $\mathbb{R}$ is denoted by $C_0^\omega(K \to \mathbb{R})$.

Note that usually analyticity is defined such that f is analytical in every point $\mathbf{x} \in K$, i.e. $f(\cdot - \mathbf{x})$ is analytical in zero for all $\mathbf{x} \in K$. Therefore the above definition considers analyticity as a local property and not as usual as a global property. Since global analyticity does not play an important role in this diploma thesis the term analytical is used short for analytical in zero.
It is important to distinguish between the two convergence concepts in (4.3.1) and (4.3.2). In the former the degree of the Taylor polynomial is fixed and $\mathbf{x}$ is tending to zero, while in the latter the point $\mathbf{x}$ is fixed and the degree of the Taylor polynomial is tending to infinity. To make the concept of analyticity clearer consider the following examples:

- The sine function is analytical in zero with $\delta_c\{\sin\} = \infty$.
- The function f given by $x \mapsto \frac{1}{x-1}$ is analytical in zero with $\delta_c\{f\} = 1$.
- The function $x \mapsto \begin{cases} 0, & x \le 0, \\ e^{-1/x}, & x > 0, \end{cases}$ is *not* analytical in zero, because all derivatives at zero are zero, but the function is not the zero function in any environment of the point zero (see Figure 7).

The last example makes clear that in general one can not expect that an increasing degree of the Taylor polynomial makes the approximation accuracy better. The aim of the next subsection is therefore to find condition under which there is a relation between the degree of the Taylor polynomial and the approximation accuracy.

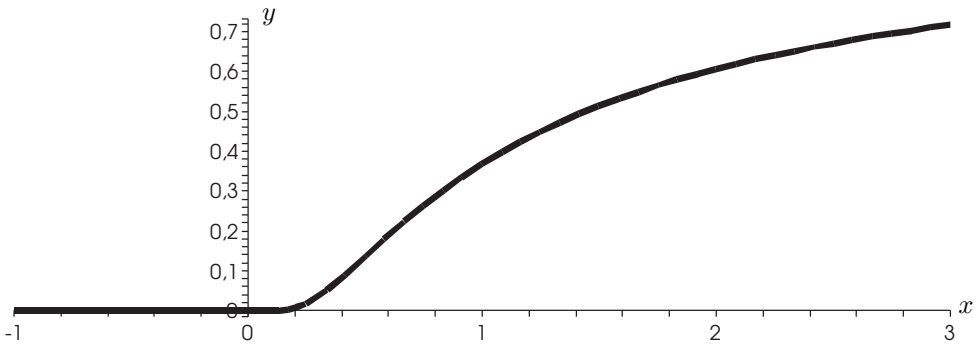

Figure 7: A C^∞-function which is not analytical.

4.4 Approximation accuracy and degree of Taylor polynomials

The definition of analytical functions ensures that the Taylor polynomial converges point-wise to the original function, but in general point-wise convergence of function sequences does not guarantee overall uniform convergence. Consider for example the function sequence $(g_n)_{n\in\mathbb{N}}$ given by $g_n : (0,1) \to \mathbb{R}$ and $g_n(x) = x^n$. Clearly $g_n(x) \to 0$ as $n \to \infty$ for every $x \in (0,1)$. On the other hand it is not possible to find for any small $\varepsilon > 0$ a $N \in \mathbb{N}$, such that $\|g_n - 0\| < \varepsilon$ for all $n \geq N$, consider for example $x_n := \sqrt[n]{0.5} \in (0,1)$, then $g_n(x_n) = 0.5$ for all $n \in \mathbb{N}$. The problem in this example is the right boundary, because at the value $x = 1$ the convergence of $g_n(x)$ to zero would not be given. In general some restrictions on the function sequence most be assumed, for the analytical functions these are given in the following definition.

Definition 4.4.1 (Nicely analytical functions)**.** *An analytical function $f \in C_0^\omega(K \to \mathbb{R})$ is called* nicely analytical *if, and only if, firstly the Taylor series converges on the whole of $K = [-1,1]^{n_0}$, i.e. the radius of convergence $\delta_c\{f\}$ fulfills $\delta_c\{f\} > 1$, and, secondly,*

$$\limsup_{k\to\infty} \sqrt[k]{\frac{\|D^k f(0)\|}{k!}} < 1. \tag{4.4.1}$$

The space of nicely analytical functions is denoted by $C^\omega_{nice}(K \to \mathbb{R})$.

Note that for $n_0 = 1$ the condition on the radius of convergence already implies (4.4.1), see, e.g., [Amann & Escher 2001*a*, Thm. 1.8], but it is not clear if this implication also holds for $n_0 > 1$. Note furthermore that by the root criteria ([Amann &

Escher 2001*a*, Thm. II.8.5]) condition (4.4.1) implies

$$\sum_{k=0}^{\infty} \frac{\|D^k f(0)\|}{k!} < \infty.$$

It is easy to see that, e.g., polynomials, sine, cosine and the exponential function are nicely analytical functions.
Finally, it is now possible to formalize the intuition that a higher degree of Taylor polynomials corresponds not only to a local accuracy increase but also to a better global approximation accuracy:

Proposition 4.4.2 (Uniform convergence of Taylor polynomials)**.** *Let $f \in C^{\omega}_{nice}(K \to \mathbb{R})$ be a nicely analytical function on $K = [-1,1]^{n_0}$, then $\mathcal{T}_N\{f\}$ converges uniformly to f as $N \to \infty$, i.e. for every arbitrarily small $\varepsilon > 0$ one can find an $N_\varepsilon \in \mathbb{N}$ such that*

$$\big\|\mathcal{T}_N\{f\} - f\big\| < \varepsilon \quad \textit{for all } N \geq N_\varepsilon.$$

Proof. From the definition of $D^n f(0)$ in Subsection 4.1 it follows that

$$\big|D^n f(0)\mathbf{x}^n\big| \leq \big\|D^n f(0)\big\|\|\mathbf{x}\|^n \leq \big\|D^n f(0)\big\|$$

for all $n \in \mathbb{N}$, $\mathbf{x} \in K = [-1,1]^{n_0}$ and $\|\mathbf{x}\| = \max\big\{|x_1|, |x_2|, \dots |x_{n_0}|\big\}$. Let $\varepsilon > 0$, then there exists, because of (4.4.1), an $N_\varepsilon \in \mathbb{N}$ such that

$$\sum_{k=N_\varepsilon+1}^{\infty} \frac{1}{k!}\|D^k f(0)\| < \varepsilon.$$

From the assumption that the radius of convergence fulfills $\delta_c\{f\} > 1$ it follows for all $\mathbf{x} \in K$

$$\big|\mathcal{T}_N\{f\}(\mathbf{x}) - f(\mathbf{x})\big| = \left|\sum_{k=N+1}^{\infty} \frac{D^k f(0)}{k!}\mathbf{x}^k\right|$$

for all $N \in \mathbb{N}$ and hence

$$\big|\mathcal{T}_N\{f\}(\mathbf{x}) - f(\mathbf{x})\big| \leq \sum_{k=N+1}^{\infty} \frac{1}{k!}\|D^k f(0)\| < \varepsilon$$

for all $N \geq N_\varepsilon$. qed

So far, there was no direct connection to MLPs, therefore the next subsection will correlate the above results about analytical functions with MLPs.

4.5 MLPs as analytical functions

The following theorem is actually only a corollary of Proposition 4.4.2, but it is a corner stone of this diploma thesis.

Theorem 4.5.1 (Global Taylor approximation)**.** *Let $f \in C^\omega_{nice}(K \to \mathbb{R})$ with $K = [-1,1]^{n_0}$ and consider an MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$ with $f_{MLP} \in C^\omega_{nice}(K \to \mathbb{R})$. For every arbitrary small $\varepsilon > 0$ there exists then $N_\varepsilon \in \mathbb{N}$, such that the following implication holds:*

$$\mathcal{T}_{N_\varepsilon}\{f\} = \mathcal{T}_{N_\varepsilon}\{f_{MLP}\} \quad \Rightarrow \quad \|f - f_{MLP}\| < \varepsilon.$$

Proof. Let $\varepsilon > 0$ be given. By Proposition 4.4.2 it is possible to choose $N_1, N_2 \in \mathbb{N}$ such that $\left\|f - \mathcal{T}_{N_1}\{f\}\right\| < \varepsilon/2$ and $\left\|\mathcal{T}_{N_2}\{f_{\mathrm{MLP}}\} - f_{\mathrm{MLP}}\}\right\| < \varepsilon/2$. With the assumption $\mathcal{T}_{N_\varepsilon}\{f\} = \mathcal{T}_{N_\varepsilon}\{f_{\mathrm{MLP}}\}$ for $N_\varepsilon := \max\{N_1, N_2\}$ and the triangle inequality for norms (see appendix) the assertion of the theorem follows:

$$\begin{aligned}
\|f - f_{\mathrm{MLP}}\| &= \left\|f - \mathcal{T}_{N_\varepsilon}\{f\} + \mathcal{T}_{N_\varepsilon}\{f_{\mathrm{MLP}}\} - f_{\mathrm{MLP}}\right\| \\
&\leq \left\|f - \mathcal{T}_{N_\varepsilon}\{f\}\right\| + \left\|\mathcal{T}_{N_\varepsilon}\{f_{\mathrm{MLP}}\} - f_{\mathrm{MLP}}\right\| \\
&\leq \left\|f - \mathcal{T}_{N_1}\{f\}\right\| + \left\|\mathcal{T}_{N_2}\{f_{\mathrm{MLP}}\} - f_{\mathrm{MLP}}\right\| \\
&< \varepsilon/2 + \varepsilon/2 = \varepsilon
\end{aligned}$$

qed

The restriction in the above theorem to nicely analytical functions f is not a hard restriction, because as already mentioned all polynomials are nicely analytical functions and hence, by Proposition 3.1.2, $C^\omega_{\mathrm{nice}}(K \to \mathbb{R})$ is dense in $C(K \to \mathbb{R})$.
The assumption that the MLP function f_{MLP} is a nicely analytical function is more critical, because this imposes restrictions to the activation function σ. The next proposition gives a sufficient condition for the activation function for which the MLP function is a nicely analytical function.

Proposition 4.5.2 (Condition on activation function)**.** *Suppose that the activation*

function $\sigma : \mathbb{R} \to \mathbb{R}$ is analytical in zero with an infinite radius of convergence, i.e.

$$\sigma(t) = \sum_{k=0}^{\infty} a_k t^k \quad \text{for all } t \in \mathbb{R},$$

for some $a_k \in \mathbb{R}$, $k \in \mathbb{N}$.
Then the MLP function f_{MLP} is nicely analytical.

For the proof of this proposition a lemma is needed:

Lemma 4.5.3 (Activation function and nicely analytical functions)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a activation function which fulfills the assumption from Proposition 4.5.2 and let $f_1, f_2, \ldots, f_m \in C^{\omega}_{nice}(K \to \mathbb{R})$ for some $m \in \mathbb{N}$ and $K = [-1, 1]^{n_0}$. The function*

$$g : K \to \mathbb{R}, \quad \mathbf{x} \mapsto \sigma\big(w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + \ldots + w_m f_m(\mathbf{x}) + \theta\big),$$

where $w_1, w_2, \ldots, w_m, \theta \in \mathbb{R}$, is then also nicely analytical.

The proof is technically and is therefore put in the appendix.

Proof of Proposition 4.5.2. Consider Definition 2.3.1 for the MLP function, then it follows inductively by Lemma 4.5.3 that the mappings $\mathbf{x} \mapsto z^{i,j}$ for each $1 \leq i \leq h$ and $1 \leq j \leq n_i$ are nicely analytical and hence the MLP function $\mathbf{x} \mapsto y$ is also nicely analytical, because the output activation function is the identity function and fulfills therefore the assumptions of Lemma 4.5.3. qed

Since the transfer function

$$\sigma : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \frac{1}{1 + e^{-t}} \tag{4.5.1}$$

is widely used in MLPs, it is studied in more detail. It is easy to see that

$$\sigma' = \sigma - \sigma^2.$$

It is also possible to express higher derivatives $\sigma^{(n)}$, $n \in \mathbb{N}$, in terms of the original function σ:

Lemma 4.5.4 (Derivatives of sigmoid activation function)**.** *The n-th derivative* $\sigma^{(n)}$ *of* σ *as in* (4.5.1) *is*

$$\sigma^{(n)} = \sum_{i=1}^{n+1} a_{i,n}\sigma^i,$$

where

$$a_{i,n} = \sum_{k=1}^{i}(-1)^{k+1}\binom{i-1}{k-1}k^n.$$

The proof is technically and is therefore put in the appendix. There, a table of the first derivatives of σ is given, too. Since the exponential function is analytical in zero, σ is also analytical in zero and

$$\sigma(t) = \sum_{i=0}^{\infty}\frac{\sigma^{(i)}(0)}{i!}t^i$$

for sufficiently small $t \in \mathbb{R}$, but to fulfill the condition of Proposition 4.5.2 this equality must hold for all $t \in \mathbb{R}$. In particular, the power series must converge for all $t \in \mathbb{R}$, i.e. the radius of convergence of the power series must be infinity. The latter is equivalent to the condition (formula of Hadamard, see, e.g., [Amann & Escher 2001*a*, Thm. 9.2])

$$\limsup_{k\to\infty}\sqrt[k]{\frac{\sigma^{(i)}(0)}{i!}} = 0.$$

Unfortunately this condition is not fulfilled, as can be seen in Figure 8. The radius of convergence can be deduced from Figure 8, and is approximately $1/0.32 \approx 3$, which can also be seen in Figure 9 where Taylor polynomials of degree 7, 19 and 99 are plotted. Clearly, an approximation outside the convergence radius is not achieved even for high degrees of the Taylor polynomials. These numerical calculation indicates that the standard activation function (4.5.1) does not fulfill the assumptions of Proposition 4.5.2 and hence it is not clear whether the MLP function of a standard MLP is nicely analytical. It seems therefore reasonable to consider other activation functions, e.g.,

- $\sigma(t) = e^t$,
- $\sigma(t) = e^{\sin(t)}$,
- $\sigma(t) = \sin(t)$,

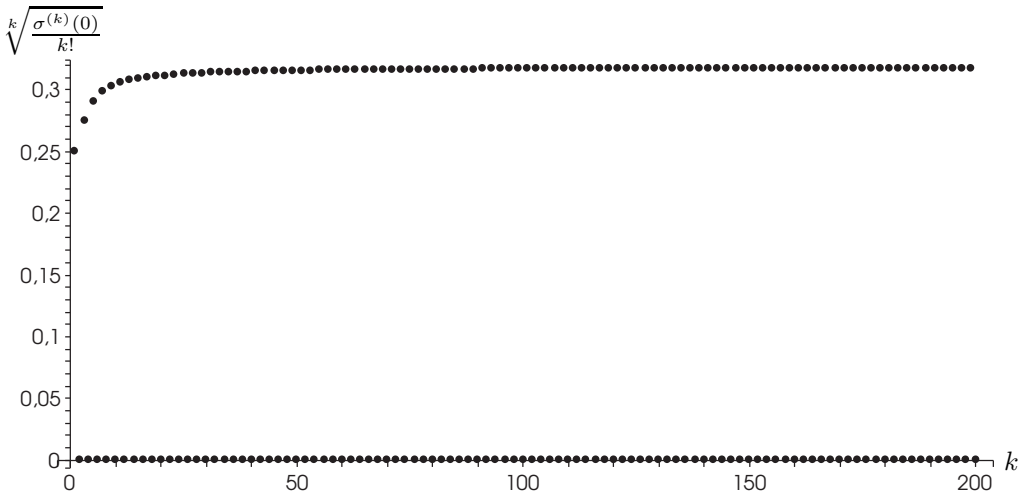

Figure 8: The k-th root of the k-th Taylor coefficient of the sigmoid activation function σ for $k = 1 \dots 200$.

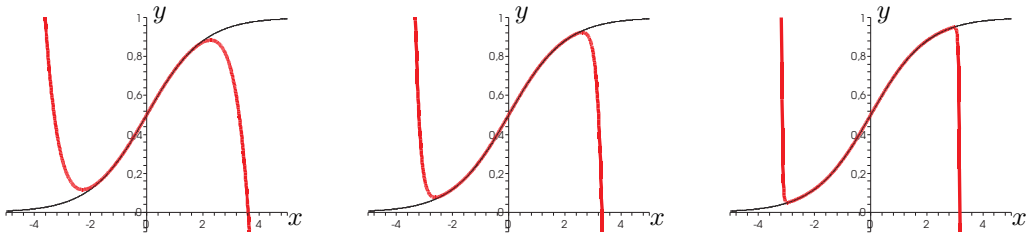

Figure 9: Taylor polynomials of degree 7, 19 and 99 for the sigmoid activation function.

which fulfill the assumptions of Proposition 4.5.2. Clearly, polynomials also fulfill the assumption of Proposition 4.5.2 and therefore the MLP function would be nicely analytical, but the condition in the implication of Theorem 4.5.1, $\mathcal{T}_{N_\varepsilon}\{f\} = \mathcal{T}_{N_\varepsilon}\{f_{\text{MLP}}\}$, is not satisfiable for an degree N_ε larger then the degree of the activation function polynomial (or multiple of this order for more than one hidden layer).
Note furthermore that one can show that for the exponential activation function the necessary condition for approximation capabilities derived in the next section are not sufficient, see Section 6 for more details.

5 Number of necessary hidden units

5.1 Main idea: Approximation order

In the previous section it was shown that under certain conditions the order of the Taylor polynomial corresponds to the approximation accuracy (Theorem 4.5.1). But even if the conditions are not fulfilled, a higher degree of the Taylor polynomial ensures a better local approximation as expressed in (4.3.1). The main idea of this diploma thesis is therefore to consider an *approximation order* instead of a global approximation accuracy:

Definition 5.1.1 (Approximation order)**.** *Consider $f, g \in C^N(K \to \mathbb{R})$ for $N \in \mathbb{N}$. The function f approximates g with order N if, and only if,*

$$\mathcal{T}_N\{f\} = \mathcal{T}_N\{g\}.$$

The question which will be answered in this diploma thesis is the following:

Given an order $N \in \mathbb{N}$, how many hidden units are necessary to ensure that an MLP can reach approximation order N for any sufficiently smooth function $f : K \to \mathbb{R}$?

For a better understanding of the following derivation of the answer to this question an example is considered in parallel.

For the example the desired approximation order is $N = 2$ and the number of inputs is $n_0 = 2$. For any $f \in C^2(K \to \mathbb{R})$,

$$\mathcal{T}_2\{f\}(\mathbf{x}) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{11}x_1x_2 + a_{20}{x_1}^2 + a_{02}{x_2}^2$$

and analogously for the MLP function

$$\mathcal{T}_2\{f_{\text{MLP}}\}(\mathbf{x}) = b_{00} + b_{10}x_1 + b_{01}x_2 + b_{11}x_1x_2 + b_{20}{x_1}^2 + b_{02}{x_2}^2.$$

The *number of coefficients* of a general Taylor polynomial of degree N with n_0 variables is denoted by

$$\mathcal{N}_{\mathcal{CT}}(N, n_0).$$

For the example

$$\mathcal{N}_{\mathcal{CT}}(2,2) = 6.$$

Furthermore, denote the vector of all coefficients of the Taylor polynomial of f with

$$\mathcal{CT}_N\{f\} \in \mathbb{R}^{\mathcal{N}_{\mathcal{CT}}(N,n_0)},$$

which is

$$\mathcal{CT}_2\{f\} = (a_{00}, a_{10}, a_{01}, a_{11}, a_{20}, a_{02}) =: \mathbf{a} \in \mathbb{R}^6$$

and

$$\mathcal{CT}_2\{f_{\mathrm{MLP}}\} = (b_{00}, b_{10}, b_{01}, b_{11}, b_{20}, b_{02}) =: \mathbf{b} \in \mathbb{R}^6$$

for the example.
Polynomials are equal if, and only if, all their coefficients are equal to each other, in particular

$$\mathcal{T}_N\{f\} = \mathcal{T}_N\{f_{\mathrm{MLP}}\} \quad \Leftrightarrow \quad \mathcal{CT}_N\{f\} = \mathcal{CT}_N\{f_{\mathrm{MLP}}\}.$$

Therefore, to ensure approximation of order N it is necessary that the latter equation is solvable. In fact, the latter equation is a system of $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ equations, for the considered example, this are six equations

$$\begin{aligned} a_{00} &= b_{00}, \\ a_{10} &= b_{10}, \\ a_{01} &= b_{01}, \\ a_{11} &= b_{11}, \\ a_{20} &= b_{20}, \\ a_{02} &= b_{02}. \end{aligned}$$

Whilst the coefficients $\mathcal{CT}_N\{f\}$ can be arbitrary, since f can be arbitrary, the coefficients $\mathcal{CT}_N\{f_{\mathrm{MLP}}\}$ are uniquely determined through the MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$. The aim of this diploma thesis is to find a minimal fixed structure, i.e. finding h, $\mathbf{n}$ and σ, such that in principle only through variation of $\mathbf{P}$ all function $f \in C^2(K \to \mathbb{R})$ can be approximated with order N. Recall that the number of parameters of a MLP $(h, \mathbf{h}, \sigma, \mathbf{P})$

with a fixed structure is denoted by

$$\mathcal{N}_{\mathrm{MLP}}(h, \mathbf{n}).$$

It is then convenient to consider the coefficients $\mathcal{CT}_N\{f_{\mathrm{MLP}}\}$ as a function of $\mathbf{P}$, i.e.

$$\mathcal{CT}_N\{f_{\mathrm{MLP}}\} = \mathcal{CT}_{N,h,\mathbf{n},\sigma}(\mathbf{P}),$$

where $\mathcal{CT}_{N,h,\mathbf{n},\sigma} : \mathbb{R}^{\mathcal{N}_{\mathrm{MLP}}(h,\mathbf{n})} \to \mathbb{R}^{\mathcal{N}_{\mathcal{CT}}(N,n_0)}$.
The question, whether it is possible to approximate any function $f \in C^N(K \to \mathbb{R})$ with order $N \in \mathbb{N}$ with an MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$, is therefore equivalent to the question, whether one can find for all $\mathbf{a} \in \mathbb{R}^{\mathcal{N}_{\mathcal{CT}}(N,n_0)}$ parameters $\mathbf{P} \in \mathbb{R}^{\mathcal{N}_{\mathrm{MLP}}(h,\mathbf{n})}$ such that

$$\mathbf{a} = \mathcal{CT}_{N,h,\mathbf{n},\sigma}(\mathbf{P}),$$

i.e. the following proposition holds.

Proposition 5.1.2 (Approximation order and coefficient function)**.** *An MLP with fixed structure, i.e. fixed $h \in \mathbb{N}$, $\mathbf{n} \in \mathbb{N}^{h+1}$ and sufficiently smooth $\sigma : \mathbb{R} \to \mathbb{R}$, is capable of approximating any function $f \in C^N(K \to \mathbb{R})$ with order $N \in \mathbb{N}$ if, and only if,*

$$\mathcal{CT}_{N,h,\mathbf{n},\sigma} : \mathbb{R}^{\mathcal{N}_{MLP}(h,\mathbf{n})} \to \mathbb{R}^{\mathcal{N}_{\mathcal{CT}}(N,n_0)} \quad \textit{is surjective}^{1}.$$

As an example consider the fixed structure $h = 1$, $\mathbf{n} = (2, 3)$ and σ is the standard activation function given by (4.5.1). The parameters $\mathbf{P}$ (in a structured form) are denoted by

$$\begin{aligned}
\mathbf{P} &= (\mathbf{W}^1, \mathbf{w}^y),\\
\mathbf{W}^1 &= (\mathbf{w}^{1,1}, \mathbf{w}^{1,2}, \mathbf{w}^{1,3}),\\
\mathbf{w}^{1,1} &= (w_1^{1,1}, w_2^{1,1}, \theta^{1,1}) \in \mathbb{R}^3,\\
\mathbf{w}^{1,2} &= (w_1^{1,2}, w_2^{1,2}, \theta^{1,2}) \in \mathbb{R}^3,\\
\mathbf{w}^{1,3} &= (w_1^{1,3}, w_2^{1,3}, \theta^{1,3}) \in \mathbb{R}^3,\\
\mathbf{w}^y &= (w_1^y, w_2^y, w_3^y) \in \mathbb{R}^3,
\end{aligned}$$

hence $\mathcal{N}_{\mathrm{MLP}}(h, \mathbf{n}) = 12$ and $\mathbf{P}$ can be considered as an element of $\mathbb{R}^{12}$.

[1] A function $f : X \to Y$ is called surjective, if, and only if, for every $y \in Y$ there exists $x \in X$ with $f(x) = y$, i.e. f maps X onto Y.

Writing f_{MLP} as $F_{\mathrm{MLP}}[\mathbf{P}]$ (see Subsection 2.3), the coefficient function is

$$\mathcal{CT}_{2,1,(2,3),\sigma}(\mathbf{P}) = \begin{pmatrix} b_{00}(\mathbf{P}) \\ b_{10}(\mathbf{P}) \\ b_{01}(\mathbf{P}) \\ b_{11}(\mathbf{P}) \\ b_{20}(\mathbf{P}) \\ b_{02}(\mathbf{P}) \end{pmatrix} = \begin{pmatrix} F_{\mathrm{MLP}}[\mathbf{P}](0) \\ \frac{\partial F_{\mathrm{MLP}}[\mathbf{P}]}{\partial x_1}(0) \\ \frac{\partial F_{\mathrm{MLP}}[\mathbf{P}]}{\partial x_2}(0) \\ 2\frac{\partial^2 F_{\mathrm{MLP}}[\mathbf{P}]}{\partial x_1 \partial x_2}(0) \\ \frac{\partial^2 F_{\mathrm{MLP}}[\mathbf{P}]}{\partial^2 x_1}(0) \\ \frac{\partial^2 F_{\mathrm{MLP}}[\mathbf{P}]}{\partial^2 x_2}(0) \end{pmatrix},$$

where

$$\begin{aligned} F_{\mathrm{MLP}}[\mathbf{P}](\mathbf{x}) = {} & w_1^y \, \sigma(w_1^{1,1} x_1 + w_2^{1,1} x_2 + \theta^{1,1}) \\ & + w_2^y \, \sigma(w_1^{1,2} x_1 + w_2^{1,2} x_2 + \theta^{1,2}) \\ & + w_3^y \, \sigma(w_1^{1,3} x_1 + w_2^{1,3} x_2 + \theta^{1,3}) \end{aligned}$$

and in particular

$$\begin{aligned} F_{\mathrm{MLP}}[\mathbf{P}](0) &= w_1^y \sigma(\theta^{1,1}) + w_2^y \sigma(\theta^{1,2}) + w_3^y \sigma(\theta^{1,3}), \\ \frac{\partial F_{\mathrm{MLP}}[\mathbf{P}]}{\partial x_1}(0) &= w_1^y w_1^{1,1} \sigma'(\theta^{1,1}) + w_2^y w_1^{1,2} \sigma'(\theta^{1,2}) + w_3^y w_1^{1,3} \sigma'(\theta^{1,3}), \end{aligned}$$

etc.

It remains now to show what conditions are necessary for the surjectivity of the function $\mathcal{CT}_{N,h,\mathbf{n},\sigma}$.

5.2 Necessary conditions for the solvability of systems of equations

In Proposition 5.1.2 it was shown that the capability of an MLP to approximate with a given order N is equivalent to surjectivity of a certain function. It is therefore of interest when functions can in principle be surjective and when not. A main feature of the function in question is that it is possible to vary the number of parameters, whilst the number of equations stay the same. For the example which was considered in the previous subsection the number of equations is always six, the number of parameters is

for the chosen network structure twelve, but if one takes only a single hidden unit, the number of parameters would be four. Intuitively, a system of six independent equations can only be solvable for all given values, if there are at least six "degrees of freedom" in the equations, i.e. there are at least six independent parameters. Unfortunately, this intuition is wrong:

Proposition 5.2.1 (Peano curves)**.** *There exist a continuous function $p : \mathbb{R} \to \mathbb{R}^2$ which is surjective.*

These kind of functions are called Peano curves or "space-filling curves" and are for example considered in [Sagan 1994]. One important property of these curves is that they are nowhere differentiable, in other words they only consists of corners, which is difficult to imagine.

As already mentioned in another context, the continuous function can exhibit strange properties. But if one considers differentiable function, the above intuition holds:

Theorem 5.2.2 (Surjectivity of differentiable functions)**.** *Let $U \subseteq \mathbb{R}^m$ be open and $g \in C^1(U \to \mathbb{R}^n)$, i.e. a differentiable function. If $m < n$ then $g(U) \neq \mathbb{R}^n$, i.e. g is not surjective.*

Proof. The following proof is based on [Merker 2006].
The space $\mathbb{R}^m$ is Lindelöf (i.e. every open covering has a countable subcovering, see, e.g., [Abraham, Marsden & Ratiu 1988]) and hence $U \subseteq \mathbb{R}^m$ is Lindelöf, too. The set

$$\mathcal{R}_g := \left\{ \; y \in \mathbb{R}^n \;\middle|\; \forall x \in g^{-1}(y) : g'(x) \in \mathcal{L}(\mathbb{R}^m \to \mathbb{R}^n) \text{ is surjective } \right\}$$

is by Sard's Theorem for Manifolds, [Abraham et al. 1988, Thm. 3.6.8], a countable intersection of open dense sets. Note that for $y \notin g(U)$ trivially $y \in \mathcal{R}_g$. On the other hand for every $y \in g(U)$ and $x \in g^{-1}(y)$ the linear mapping $g'(x)$ is not surjektive, because $m < n$. Hence $g(U) \cap \mathcal{R}_g = \emptyset$, i.e. $g(U) \subseteq \mathbb{R}^n \backslash \mathcal{R}_g$. The complement of a countable intersection of open dense sets is a countable union of closed sets with empty interior, hence $g(U) \subseteq M := \bigcup_{i \in \mathbb{N}} C_i$, where $C_i \subseteq \mathbb{R}^n$ are closed sets with empty interior.
Seeking a contradiction assume $g(U) = \mathbb{R}^n$, which implies that $M = \mathbb{R}^n$.
The space $\mathbb{R}^n$ is locally compact and the Baire Category Theorem, [Abraham et al. 1988, Thm. 1.7.3], yields that $\mathbb{R}^n$ is a Baire space, i.e. every countable intersection of

open and dense subsets is dense. For $i \in \mathbb{N}$ the subsets $O_i := \mathbb{R}^n \backslash C_i$ are open and dense and hence $\bigcap_{i\in\mathbb{N}} O_i$ is dense in $\mathbb{R}^n$. This yields the contradiction

$$\emptyset = \mathbb{R}^n \backslash M = \mathbb{R}^n \backslash \bigcup_{i\in\mathbb{N}} C_i = \bigcap_{i\in\mathbb{N}} O_i.$$

qed

The question is now, whether the Taylor coefficient function $\mathcal{CT}_{N,h,\mathbf{n},\sigma}$ is differentiable. This is answered with the following proposition:

Proposition 5.2.3 (Differentiability of Taylor coefficient function)**.** *If $\sigma \in C^\infty(\mathbb{R} \to \mathbb{R})$ then $\mathcal{CT}_{N,h,\mathbf{n},\sigma}$ is differentiable for all $N \in \mathbb{N}$, $h \in \mathbb{N}$ and $\mathbf{n} \in \mathbb{N}^{h+1}$.*

Proof. Writing $\mathcal{T}_N\{f_{\text{MLP}}\}(\mathbf{x}) = \sum_{|I|\le N} b_I(\mathbf{P})\mathbf{x}^I$ (see Subsection 3.1), the Taylor coefficient function $\mathcal{CT} : \mathbb{R}^{\mathcal{N}_{\text{MLP}}(h,\mathbf{n})} \to \mathbb{R}^{\mathcal{N}_{\mathcal{CT}}(N,n_0)}$ can be considered as $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ single functions, which have the values $b_I(\mathbf{P}) \in \mathbb{R}$ for some unique index $I \in \mathbb{N}^{n_0}$ with $|I| \le N$. It suffices now to show that for each $I \in \mathbb{N}^{n_0}$ with $|I| \le N$ the function $\mathbf{P} \mapsto b_I(\mathbf{P})$ is differentiable.
It is

$$b_I(\mathbf{P}) = c_I \frac{\partial^{|I|} F_{\text{MLP}}[\mathbf{P}]}{(\partial\mathbf{x})^I}(0),$$

where c_I is some multiple which results from the symmetries of the partial derivatives and

$$(\partial\mathbf{x})^I = \partial^{i_1} x_1 \partial^{i_2} x_2 \cdots \partial^{i_{n_0}} x_{n_0},$$

if $I = (i_1, i_2, \ldots, i_{n_0})$. From the definition of the MLP function (see Definition 2.3.1) and the assumption that $\sigma \in C^\infty(\mathbb{R} \to \mathbb{R})$ it follows that the MLP function $(\mathbf{P}, \mathbf{x}) \mapsto F_{\text{MLP}}[\mathbf{P}](\mathbf{x})$ is not only arbitrarily often continuously differentiable with respect to $\mathbf{x}$, but also with respect to $\mathbf{P}$. This implies that the function $(\mathbf{P}, \mathbf{x}) \mapsto F_{\text{MLP}}[\mathbf{P}](\mathbf{x})$ is arbitrarily often differentiable. In particular, there exists for every partial derivative a further partial derivative. The derivative of $\mathbf{P} \mapsto b_I(\mathbf{P})$ is simply the partial derivative (with respect to $\mathbf{P}$) of $c_I \frac{\partial^{|I|} F_{\text{MLP}}[\mathbf{P}]}{(\partial\mathbf{x})^I}(0)$, which itself is a partial derivative of the MLP function $(\mathbf{P}, \mathbf{x}) \mapsto F_{\text{MLP}}[\mathbf{P}](\mathbf{x})$ with respect to $\mathbf{x}$. Hence $\mathbf{P} \mapsto b_I(\mathbf{P})$ is differentiable, which implies differentiability of $\mathcal{CT}_{N,h,\mathbf{n},\sigma}$. qed

Note that for the differentiability of $\mathcal{CT}_{N,h,\mathbf{n},\sigma}$ only $\sigma \in C^{N+1}(\mathbb{R} \to \mathbb{R})$ is needed. The aim of this diploma thesis is to vary only the number of hidden units to achieve

a good approximation order, an activation function which is not smooth would artificially restrict the achievable order of approximation. Therefore only smooth activation functions are considered. Note furthermore that from smoothness of the activation function already smoothness of the Taylor coefficient function follows.
With the above results it is now possible to establish a quantitative relation between the approximation order and the structure of the MLP:

Corollary 5.2.4 (Necessary inequality condition)**.** *A necessary condition for an MLP* $(h, \mathbf{n}, \sigma, \mathbf{P})$ *with smooth* σ *to be capable of an approximation with order* $N \in \mathbb{N}$ *is*

$$\mathcal{N}_{MLP}(h, \mathbf{n}) \geq \mathcal{N}_{\mathcal{CT}}(N, n_0),$$

i.e. the number of parameters in the MLP must be at least the number of coefficients in the Taylor polynomial of degree N*.*

It remains now to give explicit formulas for $\mathcal{N}_{\mathrm{MLP}}(h, \mathbf{n})$ and $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ in order to calculate the number of necessary hidden units for a given approximation order.

5.3 Number of coefficients in multivariable polynomials

For the calculation of the number of coefficients in the Taylor polynomial the following lemma is very helpful:

Lemma 5.3.1 (Recursive formula for number of polynomial coefficients)**.** *The number* $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ *of coefficient of a Taylor-polynomial with* n_0 *variables and degree* N *fulfills:*

- $\mathcal{N}_{\mathcal{CT}}(N, 0) = 1$ *for all degrees* $N \in \mathbb{N}$*.*
- $\mathcal{N}_{\mathcal{CT}}(0, n_0) = 1$ *for all variable numbers* $n_0 \in \mathbb{N}$*.*
- $\mathcal{N}_{\mathcal{CT}}(N, n_0) = \mathcal{N}_{\mathcal{CT}}(N, n_0 - 1) + \mathcal{N}_{\mathcal{CT}}(N - 1, n_0)$ *for all* $n_0 > 0$ *and* $N > 0$*.*

Proof. A polynomial without variables is a constant and hence $\mathcal{N}_{\mathcal{CT}}(N, 0) = 1$ for all degrees $N \in \mathbb{N}$. A polynomial of degree zero is a constant, too, hence $\mathcal{N}_{\mathcal{CT}}(0, n_0) = 1$ for all variable numbers $n_0 \in \mathbb{N}$. Let now $n_0 > 0$, $N > 0$ and $p \in \mathcal{P}_N(\mathbb{R}^{n_0} \to \mathbb{R})$ be a polynomial of degree N with variables $x_1, x_2, \ldots, x_{n_0}$, i.e. p consists of $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ monomials, where also monomials are counted which have a zero-coefficient as long the degree is not bigger than N. The polynomial p can be written as

$$p(x_1, x_2, \ldots, x_{n_0-1}, x_n) = p_1(x_1, x_2, \ldots, x_{n-1}) + x_n p_2(x_1, x_2, \ldots, x_{n-1}, x_n),$$

where $p_1 \in \mathcal{P}_N(\mathbb{R}^{n_0-1} \to \mathbb{R})$ consists of all monomials of p where the variable x_n does not occur. The remaining monomials contain all the variable x_n and can therefore be written as $x_n p_2(x_1, x_2, \ldots, x_{n-1}, x_n)$, where $p_2 \in \mathcal{P}_{N-1}(\mathbb{R}^{n_0} \to \mathbb{R})$ has degree $N-1$. Hence, by induction, $\mathcal{N}_{\mathcal{CT}}(N, n_0) = \mathcal{N}_{\mathcal{CT}}(N, n_0 - 1) + \mathcal{N}_{\mathcal{CT}}(N-1, n_0)$. qed

In Lemma 5.3.1 the number $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ can only be calculated recursively, which is a bit unsatisfying. Fortunately, there exists also an explizit formula:

Proposition 5.3.2 (Number of polynomial coefficients)**.** *The number of coefficients in a Taylor polynomial with n_0 variables and degree N is*

$$\mathcal{N}_{\mathcal{CT}}(N, n_0) = \binom{N + n_0}{n_0}.$$

Proof. It suffices to show that the given value of $\mathcal{N}_{\mathcal{CT}}(N, n_0)$ fulfills the recursive formula from Lemma 5.3.1, i.e.

$$\binom{N + n_0}{n_0} = \binom{N - 1 + n_0}{n_0} + \binom{N + n_0 - 1}{n_0 - 1},$$

but this is a well known property of the binomial coefficients. qed

5.4 Number of parameters in MLP

The number of weights between two layers is the product of the numbers of units in each layer, because each unit of one layer is connected with all units of the other layer. For all layers except for the input and output layer the number of biases must be added, which is the same as the number of units in each layer. These observations are summarized in the following lemma:

Lemma 5.4.1 (Number of parameters)**.** *For an MLP $(h, \mathbf{n}, \sigma, \mathbf{P})$ the number of parameters $\mathbf{P}$ is*

$$\mathcal{N}_{MLP}(h, \mathbf{n}) = \sum_{i=1}^{h} (n_{i-1} + 1) n_i + n_h.$$

Actually, one is not so much interested in $\mathcal{N}_{\mathrm{MLP}}(h, \mathbf{n})$ but in

$$\mathcal{N}^*_{\mathrm{MLP}}(n_0, n) := \max_{h \in \mathbb{N}, |\mathbf{n}| = n} \mathcal{N}_{\mathrm{MLP}}(h, \mathbf{n}),$$

where $|\mathbf{n}| = n_1 + n_2 + \ldots + n_h$ and the maximum is only taken over $\mathbf{n} = (n_0, n_1, \ldots, n_h)$, where $n_1, ..., n_h > 0$ and $n_0 > 0$ is fixed. The number $\mathcal{N}^*_{\mathrm{MLP}}(n_0, n)$ is the maximum number of parameters which an MLP with n hidden units can have, regardless how the hidden units are distributed in the different hidden layers. For the calculation of $\mathcal{N}^*_{\mathrm{MLP}}(n_0, n)$ consider first the case that the number of hidden layers is fixed:

$$\mathcal{N}^h_{\mathrm{MLP}}(n_0, n) := \max_{|\mathbf{n}|=n} \mathcal{N}_P(h, \mathbf{n}).$$

Clearly, by Lemma 5.4.1,

$$\mathcal{N}^1_{\mathrm{MLP}}(n_0, n) = (n_0 + 1)n + n.$$

For $h = 2$ the n hidden units can be distributed to the two hidden layers such that n_2 units are in the second and $n_1 = n - n_2$ units are in the first hidden layer. Therefore, by Lemma 5.4.1,

$$\mathcal{N}^2_{\mathrm{MLP}}(n_0, n) = \max_{1 \le n_2 \le n-1} \big((n_0 + 1)(n - n_2) + (n - n_2 + 1)n_2 + n_2\big).$$

To calculate $\mathcal{N}^2_{\mathrm{MLP}}(n_0, n)$ consider for fixed $n, n_0 \in \mathbb{N}$ the function

$$m : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -x^2 + x(n - n_0 + 1) + (n_0 + 1)n,$$

then $\mathcal{N}^2_{\mathrm{MLP}}(n_0, n) = \max_{1 \le n_2 \le n-1} m(n_2)$. The real valued function m has a unique maximum at

$$x_{\max} = \frac{n - n_0 + 1}{2}.$$

Since m is a parabola, $\mathcal{N}^2_{\mathrm{MLP}}(n_0, n)$ is maximal for $n_2 = \lfloor \frac{n-n_0+1}{2} \rfloor$ and $n_2 = \lceil \frac{n-n_0+1}{2} \rceil$, where $\lfloor y \rfloor$ is the largest integer smaller or equal to $y \in \mathbb{R}$ and and $\lceil y \rceil$ is the smallest integer greater or equal to y. Note that the function value is the same for both points. The optimal value n_2 is only valid, if $1 \le n_2 \le n - 1$ otherwise one of the hidden layers would be empty. Under the general assumption that $n \ge 1$ and $n_0 \ge 0$ this yields for the optimal number n_2^* of hidden units in the second layer:

$$n_2^* = \max\left\{ \left\lfloor \frac{n - n_0 + 1}{2} \right\rfloor \ , \ 1 \right\}.$$

Hence

$$\mathcal{N}^2_{\mathrm{MLP}}(n_0,n) = \begin{cases} \frac{(n+n_0+3)^2}{4} - 2(n_0+1) & \text{if } n+n_0 \text{ odd and } n \geq n_0+1, \\ \frac{(n+n_0+3)^2-1}{4} - 2(n_0+1) & \text{if } n+n_0 \text{ even and } n \geq n_0+2, \\ (n_0+1)(n-1)+n-1 & \text{otherwise.} \end{cases}$$

For the first two cases the maximum is obtained for $n_2 = \left\lfloor \frac{n-n_0+1}{2} \right\rfloor$ and in the third case for $n_2 = 1$. For the latter case it is clearly better to take only one hidden layer, because with one hidden layer more parameters can be obtained. On the other hand, if n is large compared to n_0 the two hidden layer MLP will have more parameters then the single hidden layer MLP, because in the former the dependence on n is quadratic, whilst in the latter it is only linear. Evaluating now the inequality $\mathcal{N}^2_{\mathrm{MLP}}(n_0,n) \geq \mathcal{N}^1_{\mathrm{MLP}}P(n_0,n)$ yields the following lemma:

Lemma 5.4.2 (Two hidden layers vs. one hidden layer)**.**

$$\mathcal{N}^2_{MLP}(n_0,n) \geq \mathcal{N}^1_{MLP}(n_0,n) \quad \Leftrightarrow \quad \begin{cases} n \geq n_0+1+2\sqrt{n_0} & \textit{if } n+n_0 \textit{ odd}, \\ n \geq n_0+1+\sqrt{4n_0+1} & \textit{if } n+n_0 \textit{ even}, \end{cases}$$

It remains to consider the case of more than two hidden layers. It is

$$\mathcal{N}^3_{\mathrm{MLP}}(n_0,n) = \max_{n_2,n_3} \mathcal{N}_{\mathrm{MLP}}\big(3,(n_0,n-n_2-n_3,n_2,n_3)\big)$$

and by Lemma 5.4.1

$$\begin{aligned} &\mathcal{N}_{\mathrm{MLP}}\big(3,(n_0,n-n_2-n_3,n_2,n_3)\big) \\ &\quad = (n_0+1)(n-n_2-n_3) + (n-n_2-n_3+1)n_2 + (n_2+1)n_3 + n_3 \\ &\quad = (n_0+1)n + n_2(n-n_0-n_2-1) - n_3(n_0-1) \end{aligned}$$

Clearly, the value of n_3 must be chosen minimal to maximise $\mathcal{N}_{\mathrm{MLP}}\big(3,(n_0,n-n_2-n_3,n_2,n_3)\big)$, because $n_0 \geq 1$ (if $n_0 = 1$ then the value of the maximum does not depend on n_3 and it can also be chosen to be minimal to obtain the maximal value). Hence the optimal value n_3^* is

$$n_3^* = 1$$

and

$$\begin{aligned}\mathcal{N}^3_{\mathrm{MLP}}(n_0, n) &= \max_{n_2} \mathcal{N}_{\mathrm{MLP}}\big(3, (n_0, n - n_2 - 1, n_2, 1)\big) \\ &= \max_{n_2} (n_0 + 1)(n - n_2) + (n - n_2)n_2 - n_0 + 1 \\ &\leq \max_{n_2} (n_0 + 1)(n - n_2) + (n - n_2 + 1)n_2 + n_2 \\ &= \mathcal{N}^2_{\mathrm{MLP}}(n_0, n)\end{aligned}$$

Hence a two hidden layers MLP with the same number of hidden units as a three hidden layers MLP has always at least the same number of parameters and therefore three hidden layers are not needed if one aims for maximizing the number of parameters with respect to the number of hidden units. Clearly more then three hidden layers will yield an analogous result, i.e. to achieve a maximum number of parameters for a given number of hidden units only MLPs with one or two hidden layers must be considered. All results of this subsection are summarized in the following proposition:

Proposition 5.4.3 (Maximal number of parameters)**.** *The maximum number of parameters for an MLP with $n \in \mathbb{N}$ hidden units and n_0 inputs is*

$$\mathcal{N}^*_{MLP}(n) = \begin{cases} (n_0 + 2)n & \text{if } n \leq n_0 + 1 + 2\sqrt{n_0}, \\ \frac{(n+n_0+3)^2}{4} - 2(n_0 + 1) & \text{otherwise,} \end{cases}$$

if $n + n_0$ is odd, or

$$\mathcal{N}^*_{MLP}(n) = \begin{cases} (n_0 + 2)n & \text{if } n \leq n_0 + 1 + \sqrt{4n_0 + 1}, \\ \frac{(n+n_0+3)^2-1}{4} - 2(n_0 + 1) & \text{otherwise,} \end{cases}$$

if $n + n_0$ is even.

5.5 Main result

The aim of this diploma thesis is to answer the question:

Given an order $N \in \mathbb{N}$, how many hidden units are necessary to ensure that an MLP can reach approximation order N for any sufficiently smooth function $f : K \to \mathbb{R}$?

Combining the results from the previous subsections, namely Corollary 5.2.4, Proposition 5.3.2 and Proposition 5.4.3 it is now possible to answer this fundamental question. Firstly, the simpler case of a single hidden layer MLP is considered:

Theorem 5.5.1 (Number of necessary hidden units for single hidden layer MLPs)**.** *An MLP* $(h, \mathbf{n}, \sigma, \mathbf{P})$ *with* $h = 1$ *(i.e. one hidden layer),* $\mathbf{n} = (n_0, n_1) \in \mathbb{N}^2$ *and smooth* σ *can only achieve approximation order* $N \in \mathbb{N}$ *for any function* $f \in C^N(K \to \mathbb{R})$*,* $K = [-1, 1]^{n_0}$*, if*

$$n_1 \geq \frac{\binom{N+n_0}{n_0}}{n_0 + 2}$$

hidden units are used.

If there is no restriction to a single hidden layer MLP then in many cases a two hidden layer MLP is advantageous, but more than two hidden layers are not beneficial, which was proved in the previous subsection. The results are summarized in the following theorem.

Theorem 5.5.2 (Number of necessary hidden units)**.** *Let* $(h, \mathbf{n}, \sigma, \mathbf{P})$ *be an MLP with* $h \in \mathbb{N}$*,* $\mathbf{n} = (n_0, n_1, \ldots, n_h) \in \mathbb{N}^{h+1}$*,* $\sigma \in C^\infty(K \to \mathbb{R})$*,* $K = [-1, 1]^{n_0}$*, and parameters* $\mathbf{P}$*. Let* $N \in \mathbb{N}$ *be the desired approximation order. If*

$$\binom{N+n_0}{n_0} \leq (n_0 + 2)(n_0 + 1 + 2\sqrt{n_0})$$

then at least

$$\boxed{n \geq \frac{\binom{N+n_0}{n_0}}{n_0 + 2}}$$

hidden units are necessary to achieve approximation order $N \in \mathbb{N}$ *for any function* $f \in C^N(K \to \mathbb{R})$*, otherwise*

$$\boxed{n \geq 2\sqrt{\binom{N+n_0}{n_0} + 2(n_0 + 1)} - n_0 - 3}$$

hidden units are necessary.

In the first case an MLP with one hidden layer achieves the necessary number of parameters. For the second case the necessary number of parameters are obtained for an MLP with two hidden layers, where

$$n_1 = \left\lfloor \frac{n + n_0 - 1}{2} \right\rfloor,$$

$$n_2 = n - n_1 = \left\lceil \frac{n - n_0 + 1}{2} \right\rceil.$$

A table of the necessary number of hidden units is given in the appendix.

Remark 5.5.3 (Number of hidden layers)**.**

(i) It is never necessary to use more than one hidden layer, as can be seen from Theorem 5.5.1 (and also from Theorem 3.2.1), but if one uses only the minimal number of hidden units from the second case of Theorem 5.5.2 then one has to use two hidden layers to obtain the necessary number of parameters. The same stays true, if more than the minimal number of hidden units are used, but if the number of hidden units is large enough, then two hidden layers are not necessary any more (although two hidden layers would still be advantageous, because with the same number of hidden units more parameters are available, which in general will lead to better approximation results).

(ii) From the condition in Theorem 5.5.2 it follows that if only linear or quadratic approximation should be achieved, i.e. $N \leq 2$, then only one hidden layer is needed. On the other hand, if the desired approximation order is at least twelve, then two hidden layers are needed (in the sense of (i)).

Remark 5.5.4 (Growth of the number of necessary hidden units)**.** *If n_0 is fixed then the necessary number of hidden units grows polynomially in the approximation order N. Asymptotically (big O notation), it is for the single hidden hidden layer case $n = O(N^{n_0})$ and for the two hidden layer case $n = O(N^{n_0/2})$.*

In Figure 10 the necessary number of hidden units is plotted where the number of inputs is fixed to $n_0 = 1, \ldots, 5$. The polynomial growth can clearly be seen and it is also obvious that the growth for the single hidden layer case is much faster.

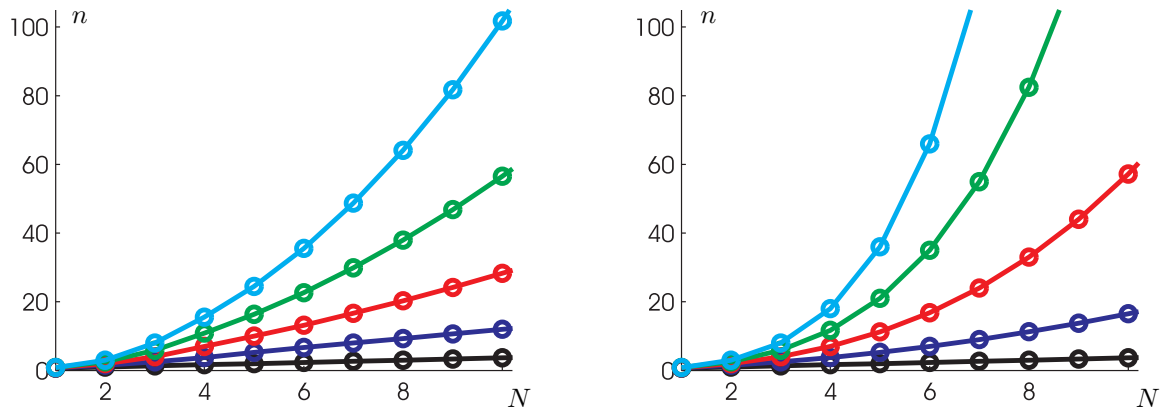

Figure 10: The number of necessary hidden units for different numbers of inputs ($n_0 = 1, \ldots, 5$ from bottom to top) without a restriction to the number of hidden layers (left picture) and with the restriction to one hidden layer (right picture).

6 Numerical simulations

6.1 MLPs with different activation functions and its Taylor polynomials

In this subsection MLPs with its Taylor polynomials are studied. For illustrative purposes only MLPs with one input are considered. As discussed in Subsection 4.5 the standard activation function has the disadvantage that its Taylor series does not converge globally, therefore other activation functions are considered as well. As activation function are considered:

- the sigmoid activation function $\sigma(t) = \frac{1}{1+e^{-t}}$,
- the exponential activation function $\sigma(t) = e^t$,
- the combined exponential-sine activation function $\sigma(t) = e^{\sin(t)}$,
- the sine activation function $\sigma(t) = \sin(t)$,

6.1.1 Taylor polynomials of MLPs $\big(1,(1,2),\sigma,\mathbf{P}\big)$

An MLP with one hidden layer and two hidden units is considered. The MLP's parameters $\mathbf{P}$ are chosen randomly, here

$$\begin{aligned}
\mathbf{P} &= (\mathbf{W}^1, \mathbf{w}^y),\\
\mathbf{W}^1 &= (\mathbf{w}^{1,1}, \mathbf{w}^{1,2}),\\
\mathbf{w}^{1,1} &= (w_1^1, \theta^{1,1}) = (-4.28829200124012, -0.45353162542848),\\
\mathbf{w}^{1,2} &= (w_1^{1,2}, \theta^{1,2}) = (-3.03409066117072, 0.35300455333831),\\
\mathbf{w}^y &= (w_1^y, w_2^y) = (0.75031647857670, -0.36420143802204).
\end{aligned}$$

Because the MLP has six parameters a maximal approximation order of five can be achieved for the class of smooth functions. The MLP function with its Taylor polynomial of degree five is plotted in Figure 11. Clearly, all Taylor polynomials are a good approximation of the MLP function in a neighbourhood of zero. The relative error is smallest for the exponential activation function, but the absolute error is smallest for the sigmoid and the sine activation and is very large for the combined exponential-sine activation function.

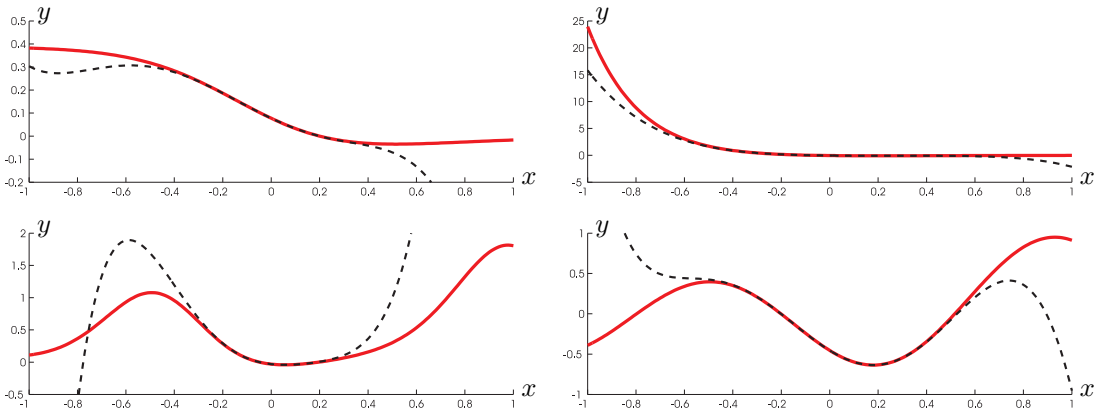

Figure 11: MLP function for two hidden units in one hidden layer and for the sigmoid, the exponential, the combined exponential-sine, and the sine activation function, resp., (solid line) and its Taylor polynomial of degree five (dashed line).

6.1.2 Taylor polynomials of MLPs $\big(1,(1,10),\sigma,\mathbf{P}\big)$

The same calculations are done again for a bigger MLP with ten hidden units (but still in one hidden layer), the parameters $\mathbf{P}$ are again chosen randomly:

$$
\begin{aligned}
\mathbf{P} &= (\mathbf{W}^1, \mathbf{w}^y),\\
\mathbf{W}^1 &= (\mathbf{w}^{1,1}, \mathbf{w}^{1,2}, \dots, \mathbf{w}^{1,10}),\\
\mathbf{w}^{1,1} &= (w_1^{1,1}, \theta^{1,1}) = (1.07523956716086, 0.68582978257388),\\
\mathbf{w}^{1,2} &= (w_1^{1,2}, \theta^{1,2}) = (0.60919569526693, 0.11531852936401),\\
\mathbf{w}^{1,3} &= (w_1^{1,3}, \theta^{1,3}) = (2.33430105798548, -0.28637475716609),\\
\mathbf{w}^{1,4} &= (w_1^{1,4}, \theta^{1,4}) = (1.15581550114654, -0.53520632660084),\\
\mathbf{w}^{1,5} &= (w_1^{1,5}, \theta^{1,5}) = (-2.12481341944397, 0.29520406952669),\\
\mathbf{w}^{1,6} &= (w_1^{1,6}, \theta^{1,6}) = (4.98395202885761, 0.98532918008095),\\
\mathbf{w}^{1,7} &= (w_1^{1,7}, \theta^{1,7}) = (-4.98822314024399, -0.23089163279376),\\
\mathbf{w}^{1,8} &= (w_1^{1,8}, \theta^{1,8}) = (0.99148596059432, -0.70432204500133),\\
\mathbf{w}^{1,9} &= (w_1^{1,9}, \theta^{1,9}) = (-3.40220869270801, 0.94327619515847),\\
\mathbf{w}^{1,10} &= (w_1^{1,10}, \theta^{1,10}) = (2.28826428716786, -0.56292896723697),\\
\mathbf{w}^{y} &= (w_1^{y}, w_2^{y}, \dots, w_{10}^{y})\\
&= (0.75031647857670, -0.36420143802204, -0.45353162542848,\\
&\quad 0.35300455333831, -0.85765840024802, -0.60681813223414,\\
&\quad 0.05815718197072, -0.65648780401238, 0.73992536400516,\\
&\quad -0.51262548758082).
\end{aligned}
$$

The number of parameters is 30, the degree of the corresponding Taylor polynomial is therefore 29. As can be seen from Figure 12 that the exponential and sine activation function yield very good approximation of the MLP function through its Taylor polynomials of degree 29 on the whole interval $[-1, 1]$. Contrarily, the sigmoid and the combined exponential-sine activation functions yield worse approximation results than for the lower order above. Of course they still achieve a good approximation in a neighbourhood of zero.

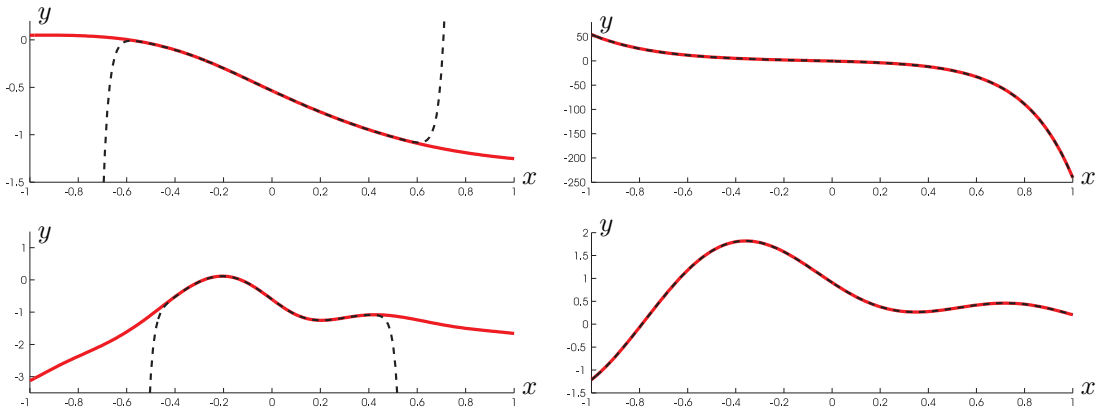

Figure 12: MLP function for ten hidden units in one hidden layer and for the sigmoid, the exponential, the combined exponential-sine, and the sine activation function, resp., (solid line) and its Taylor polynomial of degree 29 (dashed line).

It should be reminded again that these figures show only how well the MLP function is approximated by its own Taylor polynomial. Clearly, a good approximation is necessary to achieve good approximation results with other function which have the same Taylor polynomial as the MLP function. Of course, if the function which should be approximated is not well approximated through its own Taylor polynomial, then an MLP function which have this Taylor polynomial will in general not approximate the function well.

A second point is that although the exponential function seems to perform well, it has the huge drawback that, if it is used in a single hidden layer MLP, the necessary condition on the number of hidden units is definitely not sufficient to achieve a given approximation order for the whole class of smooth functions. The reason is that $\sigma(t+\theta) = \sigma(t)\sigma(\theta)$ and $\sigma' = \sigma$, which yields that the biases can always be expressed by the weights between the hidden layer and the output layer. Hence the "independent"

parameters do not include the biases and therefore more hidden units are necessary to achieve the approximation order, which is in general maximally possible for the number of parameters.

6.1.3 Taylor polynomials of MLPs $\big(2,(1,1,1),\sigma,\mathbf{P}\big)$

As in Subsection 6.1.1 two hidden units are considered, but now they are distributed in two hidden layers. This yields only five parameters and therefore the maximal approximation order is four. Note by the way that it can be seen here that it is not advantage to distribute two hidden units to two hidden layers, because the number of parameters gets lower. The randomly chosen parameters $\mathbf{P}$ are

$$\begin{aligned}
\mathbf{P} &= (\mathbf{W}^1, \mathbf{W}^2, \mathbf{w}^y),\\
\mathbf{W}^1 &= (\mathbf{w}^{1,1}),\\
\mathbf{w}^{1,1} &= (w_1^{1,1}, \theta^{1,1}) = (-2.26765812714240, -0.36420143802204),\\
\mathbf{W}^2 &= (\mathbf{w}^{2,1}),\\
\mathbf{w}^{2,1} &= (w_1^{2,1}, \theta^{2,1}) = (-4.28829200124012, 0.35300455333831),\\
\mathbf{w}^y &= (w_1^y) = (0.75031647857670).
\end{aligned}$$

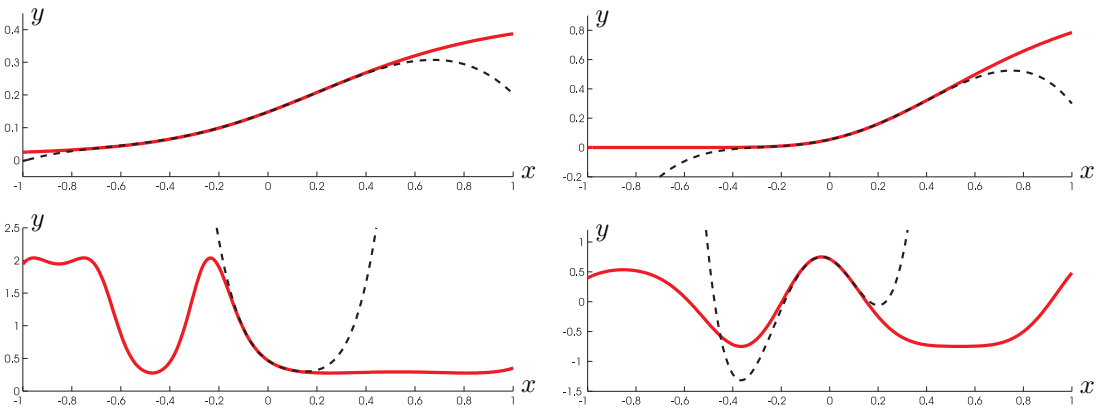

Figure 13: MLP function for one hidden unit in each of the two hidden layers and for the sigmoid, the exponential, the combined exponential-sine, and the sine activation function, resp., (solid line) and its Taylor polynomial of degree four (dashed line).

As can be seen in Figure 13 the Taylor polynomial of degree four only approximates the MLP function in a neighbourhood of zero. The best absolute and relative approxi-

mation yields the sigmoid activation function. The combined exponential-sine and sine activation yield bad approximations results.

6.1.4 Taylor polynomials of MLPs $\big(2,(1,5,5),\sigma,\mathbf{P}\big)$

Finally, a two hidden layer MLP with ten hidden units is considered, where five hidden units are in each hidden layer. This distribution of hidden units yields a total number of 45 parameters, which is much more than the number of parameters in Subsection 6.1.2, which implies that in the case of ten hidden units it is advantageous to distribute the hidden units to two hidden layers instead of one hidden layer (in agreement with Theorem 5.5.2). The 45 parameters are chosen randomly and are

$$
\begin{aligned}
\mathbf{P}&= (\mathbf{W}^1, \mathbf{W}^2, \mathbf{w}^y),\\
\mathbf{W}^1&= (\mathbf{w}^{1,1}, \mathbf{w}^{1,2}, \mathbf{w}^{1,3}, \mathbf{w}^{1,4}, \mathbf{w}^{1,5}),\\
\mathbf{w}^{1,1}&= (w_1^{1,1}, \theta^{1,1}) = (2.05748934772165, -0.60681813223414),\\
\mathbf{w}^{1,2}&= (w_1^{1,2}, \theta^{1,2}) = (0.34595558809202, 0.05815718197072),\\
\mathbf{w}^{1,3}&= (w_1^{1,3}, \theta^{1,3}) = (-0.85912427149826, -0.65648780401238),\\
\mathbf{w}^{1,4}&= (w_1^{1,4}, \theta^{1,4}) = (-1.60561897980251, 0.73992536400516),\\
\mathbf{w}^{1,5}&= (w_1^{1,5}, \theta^{1,5}) = (0.88561220858008, -0.51262548758082),\\
\mathbf{W}^2&= (\mathbf{w}^{2,1}, \mathbf{w}^{2,2}, \mathbf{w}^{2,3}, \mathbf{w}^{2,4}, \mathbf{w}^{2,5}),\\
\mathbf{w}^{2,1}&= (w_1^{2,1}, w_2^{2,1}, w_3^{2,1}, w_4^{2,1}, w_5^{2,1}, \theta^{2,1})\\
&= (0.43009582686434, 1.99358081154304, 1.24044147664523,\\
&\quad 0.40223303871561, 0.65468276770377, 0.98532918008095),\\
\mathbf{w}^{2,2}&= (w_1^{2,2}, w_2^{2,2}, w_3^{2,2}, w_4^{2,2}, w_5^{2,2}, \theta^{2,2})\\
&= (0.24367827810677, -1.99528925609760, -1.28452732408894,\\
&\quad -1.26542411452725, -1.19550998693095, -0.23089163279376),\\
\mathbf{w}^{2,3}&= (w_1^{2,3}, w_2^{2,3}, w_3^{2,3}, w_4^{2,3}, w_5^{2,3}, \theta^{2,3})\\
&= (0.93372042319419, 0.39659438423773, -0.68053257257412,\\
&\quad 0.70678113922672, -1.49664747439645, -0.70432204500133),\\
\mathbf{w}^{2,4}&= (w_1^{2,4}, w_2^{2,4}, w_3^{2,4}, w_4^{2,4}, w_5^{2,4}, \theta^{2,4})\\
&= (0.46232620045862, -1.36088347708320, -1.51582106651229,\\
&\quad -1.76278876892503, 1.07278077945675, 0.94327619515847),\\
\mathbf{w}^{2,5}&= (w_1^{2,5}, w_2^{2,5}, w_3^{2,5}, w_4^{2,5}, w_5^{2,5}, \theta^{2,5})\\
&= (-0.84992536777759, 0.91530571486715, -0.16564368704362,\\
&\quad 1.14714182472919, -1.16696599772551, -0.56292896723697),
\end{aligned}
$$

$$\begin{aligned}\mathbf{w}^y &= (w_1^y, w_2^y, w_3^y, w_4^y, w_5^y)\\ &= (0.75031647857670, -0.36420143802204, -0.45353162542848,\\ &\quad 0.35300455333831, -0.85765840024802).\end{aligned}$$

As can be seen in Figure 14 the Taylor polynomials approximate the MLP functions very well, only for the combined exponential-sine activation function the approximation is only local. The relativ error is for the sigmoid, the exponential and the sine activation function nearly the same, whilst the absolute error for the exponential activation function is large compared to the absolute error of the sigmoid and sine activation function. This is not surprising, because the MLP with exponential activation function produces very large function values. In particular, this MLP is very sensitiv to parameter changes, this will lead to difficulties in the learning process as illustrated later in Subsection 6.2.3.

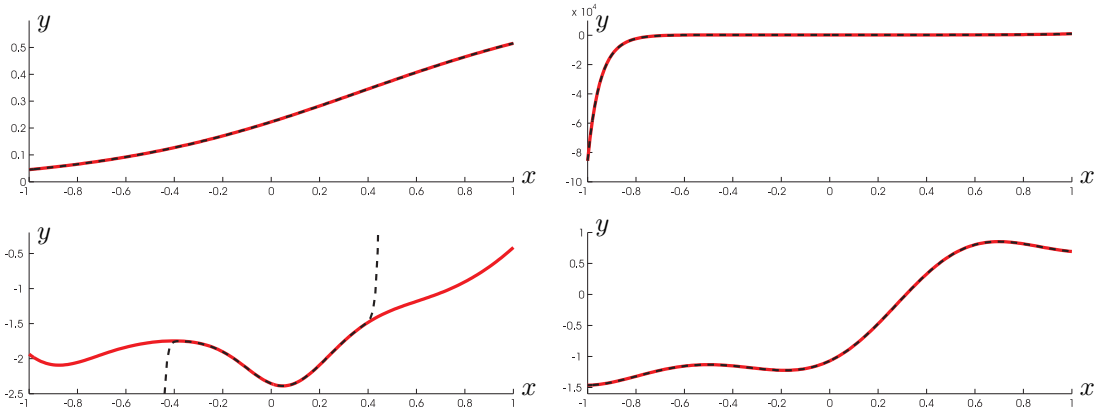

Figure 14: MLP function for five hidden units in each of the two hidden layers and for the sigmoid, the exponential, the combined exponential-sine, and the sine activation function, resp., (solid line) and its Taylor polynomial of degree 44 (dashed line).

6.2 Approximation of given polynomials with MLPs

In this subsection a given multivariable polynomial will be approximated by an MLP. The MLP will be trained with the standard back-propagation algorithm, see [Rumelhart et al. 1986] or any textbook on neural networks. There are two main issues: Firstly, the correlation between the number of hidden units and the approximation accuracy is studied and secondly, the behaviour for different activation functions is of interest.

6.2.1 Learning pattern distribution

The learning of the MLP needs learning patterns, i.e. values for the input and the corresponding outputs. Different learning patterns will yield different approximation results. To achieve global approximation accuracy the best choice are learning patterns, where the inputs are distributed uniformly in the interval of interest (i.e. $[-1, 1]$). But this will in general not give a good approximation order, because for the latter the accuracy in a neighbourhood of zero is much more important than the accuracy away from zero. Hence learning patterns where the inputs are concentrated around zero seems advantageous. To illustrate this point, a simple MLP $(1, (1, 3), \sigma, \mathbf{P})$ is considered. The maximum approximation order which this MLP can theoretically achieve is eight (the MLP has nine parameters), hence the function which should be approximated is a polynomial of degree eight with randomly chosen coefficients. In Figure 15 an MLP with the sigmoid activation function is trained for different input distributions. The dots are the 500 learning points, which are randomly chosen in

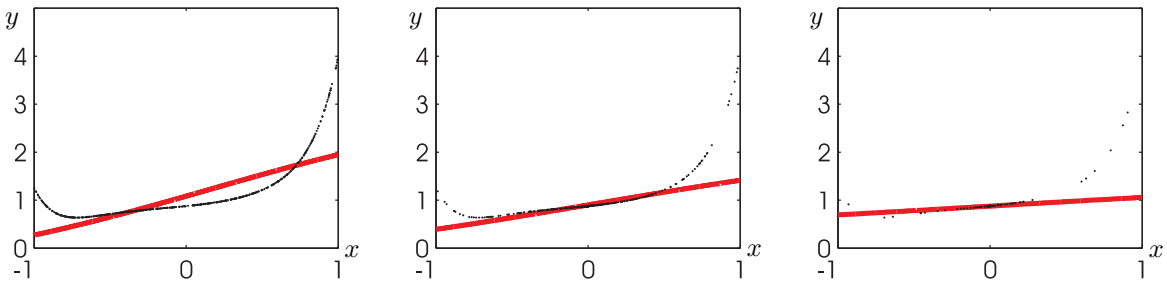

Figure 15: Sigmoid activation function: Trained MLP function (solid line) for different distribution of the sample inputs, left: uniformly distributed, middle: uniformly distributed to the power of five, right: uniformly distributed to the power of 31.

the interval $[-1, 1]$. In the left picture the distribution of the inputs x is uniform, in the middle picture the uniformly distributed input values are taken to the power of five, which yields a distribution in $[-1, 1]$ which is more concentrated around zero. In the right picture the uniformly distributed input values are taken to the power of 31, which yields an extreme concentration at zero. To judge the approximation accuracy the mean square error (MSE) is calculated. Therefore another 500 points are chosen randomly (with the same distribution as the learning pattern), the MSE for the uniformly distributed inputs is 0.179, for the inputs raised to the power of five the MSE is 0.060, and for the inputs raised to the power of 31 the MSE is 0.015.

The calculations are repeated for the other activation functions and are illustrated in Figure 16, Figure 17 and Figure 18 The mean square error for the exponential

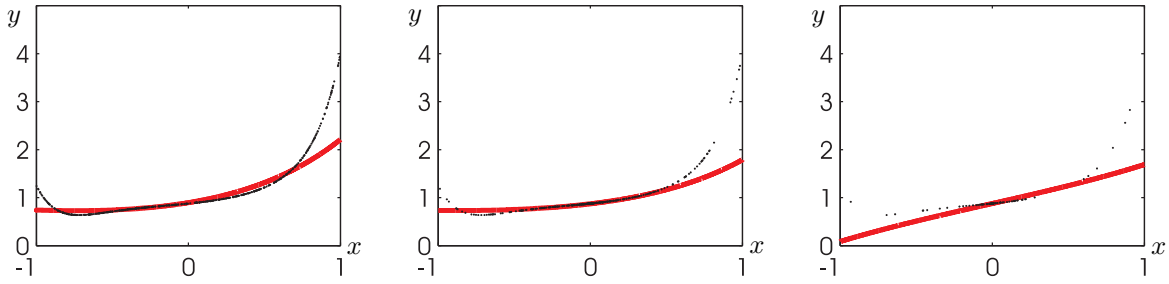

Figure 16: Exponential activation function: Trained MLP function (solid line) for different distribution of the sample inputs, left: uniformly distributed, middle: uniformly distributed to the power of five, right: uniformly distributed to the power of 31.

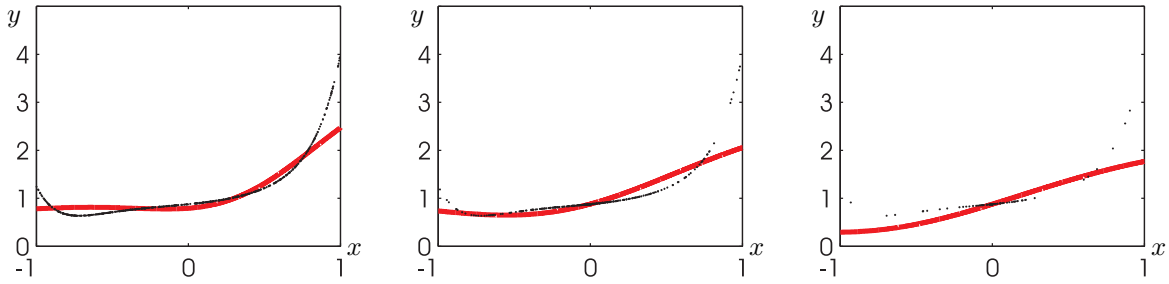

Figure 17: Combined exponential-sine activation function: Trained MLP function (solid line) for different distribution of the sample inputs, left: uniformly distributed, middle: uniformly distributed to the power of five, right: uniformly distributed to the power of 31.

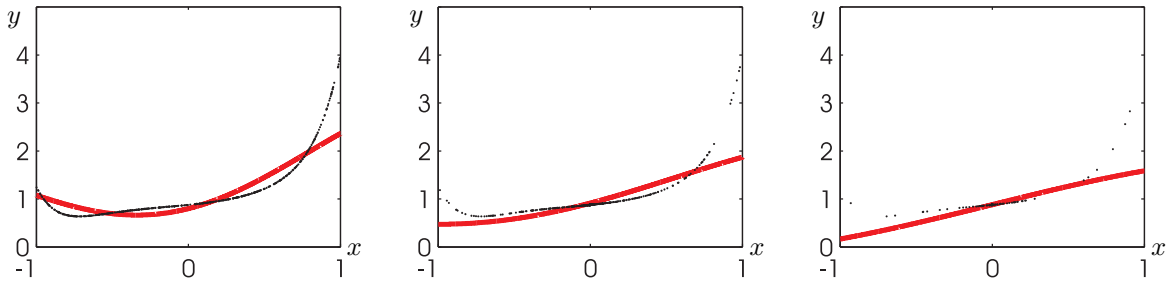

Figure 18: Sine activation function: Trained MLP function (solid line) for different distribution of the sample inputs, left: uniformly distributed, middle: uniformly distributed to the power of five, right: uniformly distributed to the power of 31.

activation function is 0.097, 0.039 and 0.008, resp., for the combined exponential-sine

activation function 0.063, 0.029 and 0.006, resp., and for the sine activation function 0.080, 0.038, and 0.008, resp.

From the Figures it can be seen that the highest power, i.e. the highest concentration around zero, does not yield the best approximation order. The reason for this might be that the most input values are so close to zero that the machine precision limits the approximation accuracy. It is an interesting question what the best distribution is for achieving the desired approximation order, but this is not in the scope of this diploma thesis. In the following numerical simulation the distribution which results from raising the uniformly distributed inputs to the power of five will be used.

6.2.2 Approximation with a single hidden layer MLP

An MLP $\big(1, (3, 60), \sigma, \mathbf{P}\big)$ is considered, the number of hidden units fulfill the necessary condition to achieve approximation order of ten (indeed for this approximation order only 58 hidden units are necessary, but to be comparable to the next subsection 60 hidden units where chosen).

The same polynomial of degree ten was considered for all four different activation functions. Each MLP was trained five times with different initial values for the network parameters (which were chosen randomly, but are the same for each activation function). For the learning 10000 different points were randomly generated (again the same for each activation function) and every 1000 learning steps the error was calculated with 1000 random points which the network had not learned yet. The distribution of the randomly generated input points is in each component the uniform distribution on $[-1, 1]$ taken to the power of five. The mean square error (in decade logarithmic scale) is plotted in Figure 19.

As can be seen from Figure 19, the sigmoid activation function compares badly to the other three activation function, the remaining error is about a factor of ten larger. Since the same learning algorithm was used, the reason for this bad performance might be the "bad" properties of the sigmoid function as described in Subsection 4.5, namely that the sigmoid function is analytical, but its Taylor series doesn't converge globally.

The same calculations are repeated for a single hidden layer MLP with 30 and 120 hidden units and are illustrated in Figure 20. In both cases the learning error is not significantly different from the error of the MLP with 60 hidden units. There are some possible reasons for this behaviour. Firstly, it is not clear if the chosen

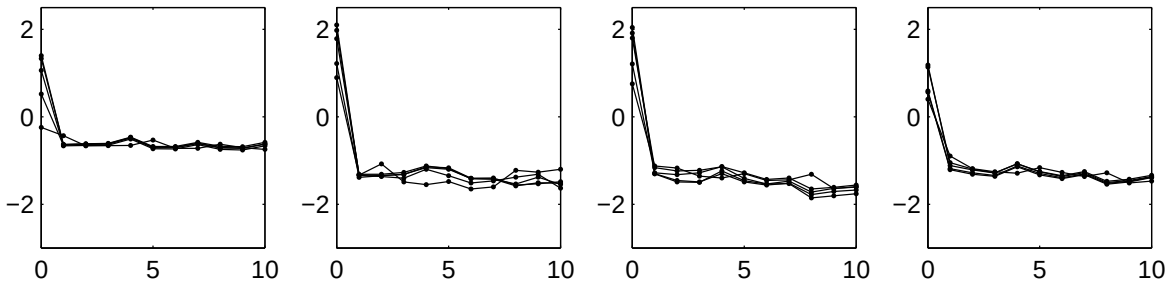

Figure 19: The mean square error (MSE) for learning a single hidden layer MLP with 60 hidden units with the sigmoid, exponential, combined exponential-sine, and sine activation function, resp., the abscissa stands for 1000 learning steps and the ordinate is the decade logarithm of the MSE.

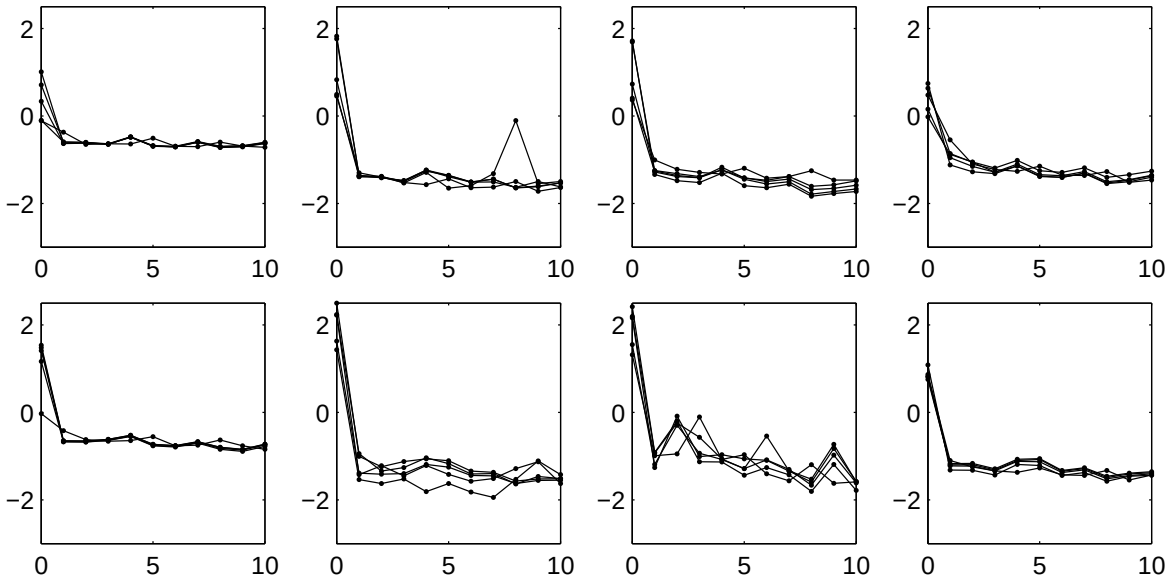

Figure 20: The mean square error (MSE) for learning a single hidden layer MLP with 30 (top four pictures) and 120 (bottom four pictures) hidden units each with the sigmoid, exponential, combined exponential-sine, and sine activation function, resp., the abscissa stands for 1000 learning steps and the ordinate is the decade logarithm of the MSE.

learning method actually achieves a good approximation order. If for example the learning method (in particular the chosen distribution of of learning inputs) can only achieve a lower approximation order than the theoretical maximally possible order for the MLP, then an MLP with a lower number of hidden units might have the same approximation accuracy. After all, the distribution for the learning inputs was only chosen heuristically, a theoretical analysis of the best distribution is an interesting topic for further research. Secondly, a better approximation order does in general not imply a better approximation accuracy, as can for example be seen in Figure 11 and Figure 12,

where for the combined exponential-sine activation function the overall accuracy is worse for a higher approximation order. At least one can see from the simulations that a much larger number of hidden units did not improve the approximation accuracy and hence more than 60 hidden units seems not to be necessary.

6.2.3 Approximation with a two hidden layers MLP

Analogue calculations as in the previous subsection are done for an MLP with two hidden layers. The number of hidden units are chosen in such a way that the two hidden layers MLP has nearly the same number of parameters as the single hidden layer MLP. A single hidden layer MLP with 60 hidden units has 300 different parameters, an MLP with two hidden layers and 15 units in the first and 14 units in the second layer has 298 different parameters. Therefore an MLP $\big(2, (1, 15, 14), \sigma, \mathbf{P}\big)$ is considered. The results of the back-propagation algorithms are shown in Figure 21. It is important to note that the learning process of the MLP with the exponentional activation function was very unstable, therefore the learning rate for this MLP is much lower than the learning rate for the other activation functions. Compared with the results for the single hidden layer

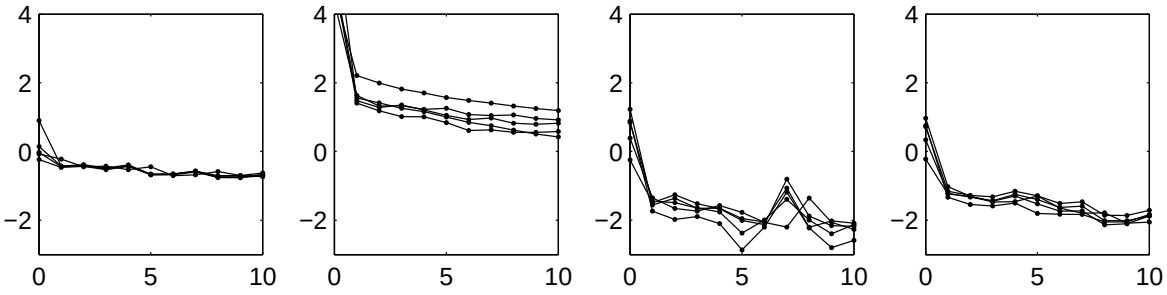

Figure 21: The mean square error (MSE) for learning an MLP with 15 units in the first hidden layer and 14 units in the second hidden layer with the sigmoid, exponential, combined exponential-sine, and sine activation function, resp., the abscissa stands for 1000 learning steps and the ordinate is the decade logarithm of the MSE.

MLP, the combined exponential-sine and sine activation functions perform significantly better. This is interesting because both networks, the single hidden layer MLP and the two hidden layers MLP, had nearly the same number of parameters (300 and 298). There seems to be no improvement for the sigmoid activation function if the hidden units are distributed in two layers instead of one and the number of parameters stays unchanged.

The learning process is also done for a smaller MLP $\big(2, (3, 8, 7), \sigma, \mathbf{P}\big)$ and a bigger MLP $\big(2, (3, 30, 28), \sigma, \mathbf{P}\big)$, see Figure 22. Again it was necessary to have a lower learning

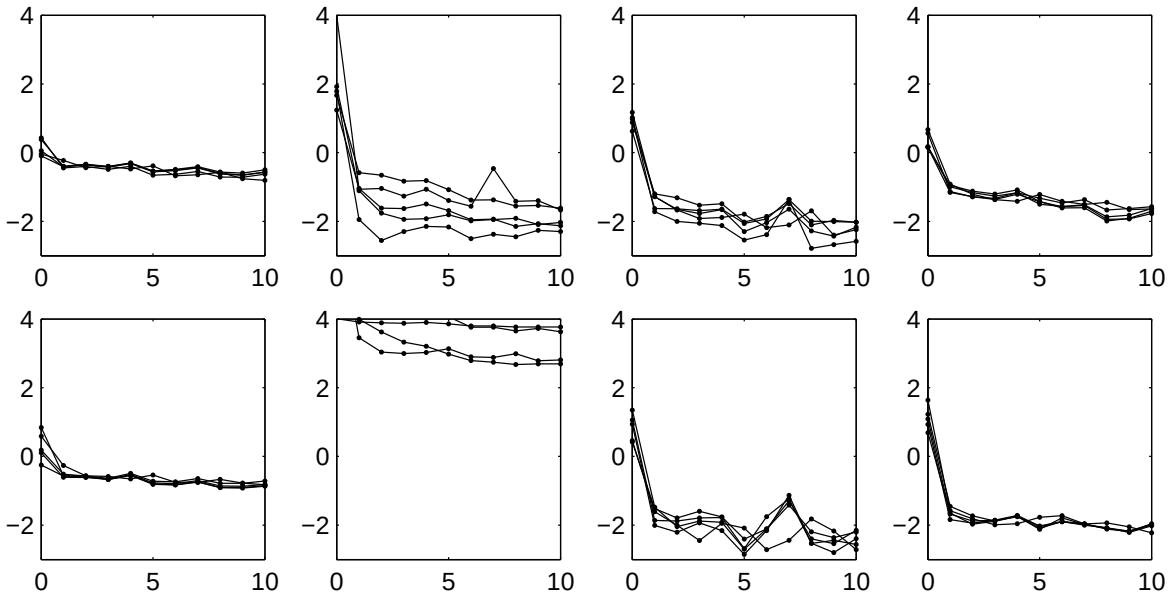

Figure 22: The mean square error (MSE) for learning an MLP with 8 units in the first hidden layer and 7 units in the second hidden layer (top four pictures) and with 30 units in the first hidden layer and 28 in the second hidden layer (bottom four pictures) each with the sigmoid, exponential, combined exponential-sine, and sine activation function, resp., the abscissa stands for 1000 learning steps and the ordinate is the decade logarithm of the MSE.

rate for the exponential activation function than for the other activation functions. The low learning rate yields that 10000 learning steps are not sufficient for large networks. It should be possible to fine tune the learning procedure to achieve faster results, but this is out of the scope of this diploma thesis. These simple simulations should only illustrate the qualitative behaviour for different activation functions and different network sizes.

7 Conclusions

The main contribution of this diploma thesis is the explicit formula for the necessary number of hidden units in a multilayer perceptron to achieve a given approximation order. It was also possible to decide how many hidden layers should be used. It turns out that more than two hidden layers are not needed, if one aims to minimize the number of necessary hidden units. Depending on the number of inputs and the desired approximation order one or two hidden layers should be used. For high approximation orders (≥ 12) two hidden layers should be used instead of one hidden layer, the same is true for smaller approximation order and a sufficiently high number of inputs, as long as the approximation order is at least three. Interestingly, for linear and quadratic approximation only one hidden layer is needed.

The correlation between approximation order and approximation accuracy was studied in detail. A sufficient condition was given for the activation function for which a high approximation order is equivalent to a high approximation accuracy. It turned out that the standard sigmoid activation function does not fulfill this condition, therefore, in the numerical simulations also other activation functions were studied. Indeed the sigmoid activation function was outperformed by the other activation functions in many situations. It seems that the sine activation function performs best for approximating functions with a given approximation order.

Although the important question "How many hidden units are necessary?" was answered in a satisfying manner, there are other important questions which remain open. The next obvious question considers the sufficient number of hidden units and under which conditions the number of necessary hidden units, calculated in this diploma thesis, is also sufficient. For the exponential activation function it was already observed in this diploma thesis that the necessary number of hidden units is not sufficient. Another important question, which was mentioned in this diploma thesis but not studied rigorously, is how an MLP must be trained to achieve a good approximation order.

8 Acknowledgements

I would like to thank my supervisor Dr. Otto who originally had the idea to correlate the approximation order with the number of hidden units in a multilayer perceptron and who let me work on this fruitful topic.

Appendix

A Tables of necessary hidden units

order	number of inputs										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	1 1	2 2	2 2	3 3	3 3	4 4	4 4	5 5	5 5	6 6	6 6
3	2 2	3 3	4 4	6 6	8 8	11 11	14 (10,4)	17 (12,5)	20 (13,6)	24 (16,7)	28 (18,8)
4	2 2	4 4	7 7	12 (7,4)	18 (10,6)	27 (13,8)	37 (17,11)	50 (21,14)	65 (25,18)	84 (30,21)	105 (35,26)
5	2 2	6 6	12 (6,4)	21 (10,7)	36 (14,11)	58 (20,15)	88 (26,21)	129 (34,28)	182 (43,35)	251 (53,44)	336 (64,55)
6	3 3	7 (4,3)	17 (8,6)	35 (13,10)	66 (20,16)	116 (29,24)	191 (40,34)	301 (53,46)	455 (69,61)	668 (88,79)	952 (109,100)
7	3 3	9 (4,4)	24 (9,8)	55 (16,14)	114 (26,23)	215 (40,35)	382 (57,51)	644 (78,72)	1040 (105,98)	1621 (138,129)	2448 (176,167)
8	3 3	12 (5,5)	33 (11,10)	83 (20,18)	184 (34,31)	376 (53,48)	715 (78,73)	1287 (112,105)	2210 (154,146)	3647 (207,199)	5814 (273,263)
9	4 4	14 (6,5)	44 (13,12)	120 (25,22)	286 (43,39)	626 (69,64)	1272 (105,100)	2431 (154,147)	4420 (219,211)	7699 (302,293)	12920 (408,398)

Table 1: Necessary number of hidden units for an MLP with 1 to 11 inputs and desired approximation order 1 to 9. The first entry is the necessary number of hidden units for an MLP with one hidden layer, the second entry is the necessary number of hidden units for a two hidden layer MLP, if the second entry consists only of one number, then a second layer is not necessary.

order	number of inputs							
	12	**13**	**14**	**15**	**16**	**17**	**18**	**19**
1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	7 7	7 7	8 8	8 8	9 9	9 9	10 10	10 10
3	33 (20,9)	38 (22,11)	43 (25,12)	48 (27,14)	54 (30,15)	60 (32,17)	67 (35,18)	74 (38,20)
4	130 (41,30)	159 (47,36)	192 (54,41)	228 (61,47)	270 (68,53)	315 (76,60)	366 (84,67)	422 (92,75)
5	442 (77,66)	572 (91,79)	727 (106,93)	912 (123,109)	1131 (141,126)	1386 (160,145)	1683 (182,165)	2024 (204,187)
6	1326 (134,124)	1809 (163,151)	2423 (195,182)	3192 (231,218)	4146 (271,257)	5313 (316,300)	6730 (365,348)	8434 (419,401)
7	3600 (223,212)	5168 (276,265)	7268 (339,327)	10032 (411,398)	13620 (493,479)	18216 (586,571)	24035 (691,675)	31324 (809,792)
8	8998 (353,342)	13566 (449,438)	19986 (564,551)	28842 (698,685)	40860 (856,841)	56925 (1038,1023)	78114 (1248,1231)	105718 (1488,1471)
9	20995 (540,530)	33162 (703,692)	51075 (902,890)	76912 (1141,1128)	113499 (1427,1413)	164450 (1766,1750)	234342 (2163,2146)	328900 (2626,2609)

Table 2: Necessary number of hidden units for an MLP with 12 to 19 inputs and desired approximation order 1 to 9. The first entry is the necessary number of hidden units for an MLP with one hidden layer, the second entry is the necessary number of hidden units for a two hidden layer MLP, if the second entry consists only of one number, then a second layer is not necessary.

order	number of inputs									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	4 4	17 (6,6)	58 (15,14)	167 (30,27)	429 (53,49)	1001 (88,83)	2161 (138,132)	4376 (207,201)	8398 (302,294)	15397 (428,419)
11	4 4	20 (7,7)	73 (17,16)	228 (35,33)	624 (64,61)	1547 (109,105)	3536 (176,171)	7559 (273,266)	15270 (408,400)	29393 (592,583)
12	5 (2,3)	23 (8,7)	91 (20,18)	304 (41,38)	884 (77,73)	2321 (134,130)	5599 (223,217)	12597 (353,346)	26721 (540,533)	53888 (802,794)
13	5 (2,3)	27 (9,8)	112 (22,20)	397 (47,44)	1224 (91,87)	3392 (163,158)	8614 (276,271)	20349 (449,443)	45220 (703,696)	95339 (1068,1059)
14	5 (2,3)	30 (9,9)	136 (24,23)	510 (53,51)	1662 (106,102)	4845 (195,190)	12920 (339,334)	31977 (563,557)	74290 (902,894)	163438 (1398,1390)
15	6 (2,3)	34 (10,9)	164 (27,25)	646 (60,58)	2215 (123,119)	6783 (231,226)	18950 (411,405)	49032 (698,692)	118864 (1141,1134)	272397 (1806,1797)
16	6 (3,3)	39 (11,10)	194 (29,28)	808 (68,65)	2907 (141,137)	9327 (271,267)	27240 (493,488)	73548 (856,849)	185725 (1427,1420)	442645 (2303,2294)
17	6 (3,3)	43 (11,11)	228 (32,30)	998 (75,73)	3762 (160,157)	12619 (316,311)	38456 (586,581)	108158 (1038,1031)	284050 (1766,1758)	703024 (2903,2894)
18	7 (3,3)	48 (12,11)	266 (35,33)	1220 (84,81)	4807 (181,178)	16825 (365,360)	53412 (691,686)	156228 (1248,1241)	426075 (2163,2155)	1093593 (3621,3612)

Table 3: Necessary number of hidden units for an MLP with 1 to 10 inputs and desired approximation order 10 to 18. The first entry is the necessary number of hidden units for an MLP with one hidden layer, the second entry is the necessary number of hidden units for a two hidden layer MLP, if the second entry consists only of one number, then a second layer is not necessary.

B Mathematical background and proofs

B.1 Metric and normed spaces

Definition B.1.1 (Metric)**.** *Let M be some set. A mapping $d : M \times M \to \mathbb{R}_{\geq 0}$ is called a* metric *if, and only if, for all $x, y, z \in M$*

M1: $d(x, y) = 0 \;\Leftrightarrow\; x = y$,

M2: $d(x, y) = d(y, x)$, *and*

M3: $d(x, y) \leq d(x, z) + d(z, y)$ *(triangle inequality).*

If d is a metric on M, then (M, d) is called metric space.

If the metric d is clear from the context, then sometimes M is also called metric space. Note that every set S is a metric space with the discrete metric, which is defined by $d(s, s) = 1$ for all $s \in S$ and zero otherwise, but this metric doesn't have much practical relevance.

For a metric space (M, d) the open ε-ball $\mathbb{B}_\varepsilon(m)$ at $m \in M$ for $\varepsilon > 0$ is defined as

$$\mathbb{B}_\varepsilon(m) := \{\; x \in M \;\mid d(x, m) < \varepsilon \;\}.$$

A subset $O \subseteq M$ is said to be *open* if, and only if for all $o \in O$ there exists $\varepsilon > 0$ such that $\mathbb{B}_\varepsilon(o) \subseteq O$. A subset $C \subseteq M$ is called *closed* if, and only if, its complement $M \backslash C$ is open.

Definition B.1.2 (Compactness)**.** *Let (M, d) be a metric space. A subset $K \subseteq M$ is called* compact *if, and only if, for all families $\mathcal{O} \subseteq \mathcal{P}(M)$ of open sets of M which cover K, i.e. $\bigcup \mathcal{O} \supseteq K$, there exists a finite subfamily $\{O_1, O_2, \ldots, O_N\} \subseteq \mathcal{O}$, $N \in \mathbb{N}$, which still covers K.*

A compact set is always closed and bounded, the converse is in general not true.

Definition B.1.3 (Norm)**.** *Let V be a real vector space (i.e. scalar multiplication and addition is defined). A mapping $\|\cdot\| : V \to \mathbb{R}_{\geq 0}$ is called a* norm *if, and only if, for all $v, w \in V$ and $\lambda \in \mathbb{R}$*

N1: $\|v\| = 0 \;\Leftrightarrow\; v = 0$,

N2: $\|\lambda v\| = |\lambda| \, \|v\|$*, and*

N3: $\|v + w\| \leq \|v\| + \|w\|$ *(triangle inequality).*

If $\|\cdot\|$ *is a norm on* V*, then* $\big(V, \|\cdot\|\big)$ *is called* normed space.

Again, as for metric spaces, if the norm is clear from the context, then V is also called a normed space. A classical normed space is the euclidian $\mathbb{R}^n$, $n \in \mathbb{N}$, where $\|\mathbf{x}\| = \sum_{i=1}^{n} {x_i}^2$ for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. For $\mathbb{R}^n$ there are also other norms possible, e.g., $\|\mathbf{x}\| = \max\big\{|x_1|, \ldots, |x_n|\big\}$, which is mostly used in this diploma thesis. It can be shown that all possible norms on $\mathbb{R}^n$ are equivalent, i.e. they are essentially the same and most properties, like continuity, convergence, boundedness, do not depend on the specific chosen norm.

Every normed space $\big(V, \|\cdot\|\big)$ is also a metric space with the induced metric $d : V \times V \to \mathbb{R}_{\geq 0}$ given by

$$d(v, w) = \|v - w\|,$$

which can easily be shown to fulfill the properties M1 - M3. Hence openness and compactness can also be considered in normed spaces. It is a well known classical result (see, e.g., [Amann & Escher 2001*a*, Thm. III.3.5]) that a subset $K \subseteq \mathbb{R}^n$ is compact if, and only if, K is closed and bounded.

The space of continuous function $C(K \to \mathbb{R})$ on a compact set $K \subseteq \mathbb{R}^n$ is a normed space:

Proposition B.1.4 ($C(K \to \mathbb{R})$ as normed space)**.** *The mapping* $\|\cdot\| : C(K \to \mathbb{R}) \to \mathbb{R}_{\geq 0}$ *given by*

$$\|f\| = \max_{x \in K} \big|f(x)\big| < \infty$$

is a norm on $C(K \to \mathbb{R})$*.*

Proof. The space $C(K \to \mathbb{R})$ is a Vector space with scalar multiplication and addition (pointwise). Every continuous function is bounded on a compact set and the maximum is attained in K (see, e.g., [Amann & Escher 2001*a*, Kor. III.3.8]) and therefore $\|f\|$ is well defined for all $f \in C(K \to \mathbb{R})$. Clearly N1 is fulfilled, because if $f \equiv 0$, then $\|f\| = 0$ and if $\|f\| = 0$ it must be $\big|f(x)\big| = 0$ for all $x \in K$, hence f is the zero function. N2 is also fulfilled, because

$$\|\lambda f\| = \max_{x \in K} \big|\lambda f(x)\big| = \max_{x \in K} |\lambda| \big|f(x)\big| = |\lambda| \|f\|.$$

Finally, N3 is fulfilled as well, because

$$\|f+g\| = \max_{x\in K} \big|f(x)+g(x)\big| \le \max_{x\in K} \big(\big|f(x)\big| + \big|g(x)\big|\big)$$
$$\le \max_{x\in K} \big|f(x)\big| + \max_{x\in K} \big|g(x)\big| = \|f\| + \|g\|.$$

qed

B.2 Banach spaces

Banach spaces are *complete* normed spaces. Completeness means, loosely speaking, that there are no gaps, e.g., the space of rational numbers is not complete, because, for example, the number $\sqrt{2}$ is not rational. The real numbers are complete. For a precise definition of completeness the concept of *Cauchy sequences* is needed:

Definition B.2.1 (Cauchy sequence)**.** *Let $\big(V, \|\cdot\|\big)$ be a normed space and let $(v_n)_{n\in\mathbb{N}}$ be a sequence in V. The sequence (v_n) is called* Cauchy sequence *if, and only if, for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ with*

$$\|v_n - v_m\| < \varepsilon \quad \forall\ n, m \ge N.$$

Note that in general a Cauchy sequence need not to converge, for example the rational sequence (v_n) given by $v_n = (1+1/n)^n$ is a Cauchy sequence, but it does not converge in the space of rational numbers, because the limit in $\mathbb{R}$ is Euler's number e, which is not a rational number. On the other hand the rational sequence (v_n) given by $v_n = 1/n$ is also a Cauchy sequence and does converge in $\mathbb{Q}$.

Definition B.2.2 (Banach space)**.** *A normed space $\big(V, \|\cdot\|\big)$ is called* complete *or* Banach space *if, and only if, all Cauchy sequences have a limit in V.*

Note that the concept of completeness can also be defined for metric spaces, but a complete metric space which is not a Banach space does not play any role here. As mentioned above all norms on $\mathbb{R}^n$ are equivalent, in particular, the normed space $\mathbb{R}^n$ is always a Banach space, regardless which norm is considered. It was already shown that the space of continuous functions $C(K \to \mathbb{R})$ on some compact K is a normed space. The following proposition gives a statement about the completeness of this normed space:

Proposition B.2.3 (Completeness of $C(K \to \mathbb{R})$)**.** *Let $K \subseteq \mathbb{R}^n$ be some compact set.*

(i) The normed space $\big(C(K \to \mathbb{R}), \|\cdot\|\big)$ with the maximum norm $\|\cdot\|$ as defined in Proposition B.1.4 is a Banach space.

(ii) Define for $p \geq 1$ the $p-$norm

$$\|\cdot\|_p : C(K \to \mathbb{R}) \to \mathbb{R}_{\geq 0}, \quad f \mapsto \|f\| := \left(\int_K \big|f(x)\big|^p dx\right)^{1/p},$$

then $\big(C(K \to \mathbb{R}), \|\cdot\|_p\big)$ is a normed space, but not a Banach space.

Proof. The proof of the first assertion can be found in [Amann & Escher 2001*a*, Thm. 2.6]. The properties N1 and N2 for $\|\cdot\|_p$ are easy to see, while N3 is well known as Minkowski's inequality. It remains to show that $\big(C(K \to \mathbb{R}), \|\cdot\|_p\big)$ is not complete, i.e. that there exist a Cauchy sequence in $C(K \to \mathbb{R})$ which does not converge in $C(K \to \mathbb{R})$. Without restriction let $K = [0, 2]$ and consider the function sequence $(f_n)_{n\in\mathbb{N}}$ given by

$$f_n : [0, 2] \to \mathbb{R}, \quad x \mapsto \begin{cases} x^n, & x \in [0, 1] \\ 1, & x > 1 \end{cases}$$

Clearly f_n is continuous for every $n \in \mathbb{N}$ and, for $n \geq m$,

$$\|f_m - f_n\|_p = \int_0^1 \big|x^m(1 - x^{n-m})\big|^p \mathrm{d}x \leq \int_0^1 x^{pm} \mathrm{d}x \leq \frac{1}{mp + 1}.$$

Hence for $\varepsilon > 0$ and $N \geq \frac{1}{p\varepsilon}$ it is for all $n \geq m \geq N$

$$\|f_m - f_n\|_p < \varepsilon,$$

i.e. (f_n) is a Cauchy sequences. The function sequence converges pointwise to the function $f : [0, 2] \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0, & x \in [0, 1), \\ 1, & x \in [1, 2], \end{cases}$$

which is not a continuous function and hence (f_n) does not converge in $C(K \to \mathbb{R})$. qed

B.3 Proof of Proposition 3.1.2

B.3.1 Denseness of $\mathcal{P}(K \to \mathbb{R})$ in $C(K \to \mathbb{R})$

This proof is based on the proof of the more general result in [Amann & Escher 2001*a*, Thm. V.4.7].

Step 1: It will be shown that for every $\delta > 0$ and for every $f \in \mathcal{P}(K \to \mathbb{R})$ there exists $g \in \mathcal{P}(K \to \mathbb{R})$ with $\big\||f| - g\big\| < \delta$.

The generalized binomial coefficient $\binom{\alpha}{k}$ for $\alpha \in \mathbb{R}$ and $k \in \mathbb{N}$ is defined as

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)(\alpha-2)\cdots\big(\alpha-(k-1)\big)}{k!}$$

and $\binom{\alpha}{0} = 1$. Many of the properties for the classical binomial coefficient still holds, in particular

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k, \quad \text{for all } x \in [-1,1],$$

see, e.g., [Amann & Escher 2001*a*, Thm. V.3.10], where it is also shown that

$$\sup_{x\in[-1,1]} \left| (1+x)^\alpha - \sum_{k=0}^{n} \binom{\alpha}{k} x^k \right| \to 0$$

as $n \to \infty$, i.e. the convergence is uniformly on the whole interval $[-1,1]$. With $\alpha = 1/2$ and $x = t^2 - 1$ this yields for all $t \in [-1,1]$:

$$|t| = \sum_{k=0}^{\infty} \binom{1/2}{k} (t^2-1)^k,$$

and for every $\tilde{\delta} > 0$ there exists $N_{\tilde{\delta}} \in \mathbb{N}$ with

$$\left| |t| - \underbrace{\sum_{k=0}^{N_{\tilde{\delta}}} \binom{1/2}{k} (t^2-1)^k}_{P_{\tilde{\delta}}(t)} \right| < \tilde{\delta} \quad \text{for all } t \in [-1,1],$$

where $P_{\tilde{\delta}} \in \mathcal{P}_{N_{\tilde{\delta}}}\big([-1,1] \to \mathbb{R}\big)$.
Let $f \in C(K \to \mathbb{R})$ and $\delta > 0$. Since f is continuous and K is compact, f is bounded and hence $\|f\| < \infty$ (see Proposition B.1.4). Let $\tilde{\delta} := \delta/\|f\| > 0$ and for an arbitrary but fixed $\mathbf{x} \in K$ let $t := f(\mathbf{x})/\|f\| \in [-1,1]$, then

$$\Big|\big|f(\mathbf{x})/\|f\|\big| - P_{\tilde{\delta}}\big(f(\mathbf{x})/\|f\|\big)\Big| < \tilde{\delta} = \delta/\|f\|, \quad \text{for all } \mathbf{x} \in K.$$

With $g_\delta : K \to \mathbb{R},\ \mathbf{x} \mapsto \|f\| P_{\tilde{\delta}}\big(f(\mathbf{x})/\|f\|\big)$ it is $g_\delta \in \mathcal{P}(K \to \mathbb{R})$ and

$$\Big|\big|f(\mathbf{x})\big| - g_\delta(\mathbf{x})\Big| < \delta \quad \forall\, \mathbf{x} \in K.$$

Step 2: It will be shown that for every $f \in C(K \to \mathbb{R})$ and $\varepsilon > 0$ there exists $h \in C(K \to \mathbb{R})$ and $p \in \mathcal{P}(K \to \mathbb{R})$ with $\|f - h\| < \varepsilon/2$ and $\|h - p\| < \varepsilon/2$.
Let $h_{\mathbf{y},\mathbf{y}}(\mathbf{x}) = f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in K$ and for $\mathbf{y} \neq \mathbf{z} \in K$ let $h_{\mathbf{y},\mathbf{z}}$ be given by

$$h_{\mathbf{y},\mathbf{z}}(\mathbf{x}) = f(\mathbf{y}) + (\mathbf{y} - \mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \frac{f(\mathbf{z}) - f(\mathbf{y})}{(\mathbf{y} - \mathbf{z}) \cdot (\mathbf{y} - \mathbf{z})},$$

which is a polynomial in $\mathbf{x}$ of degree two (or less if $f(\mathbf{z}) = f(\mathbf{y})$). Note that $h_{\mathbf{y},\mathbf{z}}(\mathbf{y}) = f(\mathbf{y})$ and $h_{\mathbf{y},\mathbf{z}}(\mathbf{z}) = f(\mathbf{z})$. Define for $\mathbf{y}, \mathbf{z} \in K$

$$\begin{aligned} U_{\mathbf{y},\mathbf{z}} &:= \{\ \mathbf{x} \in K \ \mid h_{\mathbf{y},\mathbf{z}}(\mathbf{x}) < f(\mathbf{x}) + \varepsilon/2\ \}, \\ V_{\mathbf{y},\mathbf{z}} &:= \{\ \mathbf{x} \in K \ \mid h_{\mathbf{x},\mathbf{y}}(\mathbf{x}) > f(\mathbf{x}) - \varepsilon/2\ \}, \end{aligned}$$

which are open, because $h_{\mathbf{y},\mathbf{z}} - f$ is continuous. Furthermore, $\mathbf{y} \in U_{\mathbf{y},\mathbf{z}}$ and $\mathbf{z} \in V_{\mathbf{y},\mathbf{z}}$ for all $\mathbf{y}, \mathbf{z} \in K$. For any fixed $\mathbf{z} \in K$ the family $\{\ U_{\mathbf{y},\mathbf{z}} \mid \mathbf{y} \in K\ \}$ is an open covering of the compact set K, which implies that there exists $N_{\mathbf{z}} \in \mathbb{N}$ points $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_{\mathbf{z}}}$ with

$$K \subseteq \bigcup_{i=1}^{N_{\mathbf{z}}} U_{\mathbf{y}_i,\mathbf{z}}.$$

Define

$$h_{\mathbf{z}} := \min_{1 \le i \le N_{\mathbf{z}}} h_{\mathbf{y}_i,\mathbf{z}},$$

then

$$h_{\mathbf{z}}(\mathbf{x}) < f(\mathbf{x}) + \varepsilon/2.$$

From $\min\{a,b\} = \frac{1}{2}\big(a+b-|a-b|\big)$ and Step 1 it follows inductively that for every $\delta > 0$ there exists a polynomial $p_{\delta,\mathbf{z}} \in \mathcal{P}(K \to \mathbb{R})$ with $\|h_{\mathbf{z}} - p_{\delta,\mathbf{z}}\| < \delta$.
For $\mathbf{z} \in K$ let $V_{\mathbf{z}} := \bigcap_{i=1}^{N_{\mathbf{z}}} V_{\mathbf{y}_i,\mathbf{z}}$, then $V_{\mathbf{z}}$ is open and $\mathbf{z} \in V_{\mathbf{z}}$ and, therefore the family $\{\ V_{\mathbf{z}} \mid \mathbf{z} \in K\ \}$ is an open covering of K and by compactness of K there exists $N \in \mathbb{N}$ points $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ with

$$K \subseteq \bigcup_{i=1}^{N} V_{\mathbf{z}_i}.$$

Furthermore, for all $\mathbf{z} \in K$,

$$h_{\mathbf{z}}(\mathbf{x}) > f(\mathbf{x}) - \varepsilon/2 \quad \text{for all } \mathbf{x} \in V_{\mathbf{z}}.$$

Define

$$h := \max_{1\le i\le N} h_{\mathbf{z}_i},$$

then

$$\|f - h\| < \varepsilon/2.$$

It remains to show that there exists a polynomial $p \in \mathcal{P}(K \to \mathbb{R})$ with $\|h - p\| < \varepsilon/2$. Let therefore $\delta := \frac{\varepsilon}{4}$ and chose $p_{\delta,\mathbf{z}_i} \in \mathcal{P}(K \to \mathbb{R})$ such that $\|h_{\mathbf{z}_i} - p_{\delta,\mathbf{z}_i}\| < \delta$, then

$$\left\| h - \max_{1\le i\le N} p_{\delta,\mathbf{z}_i} \right\| < \delta = \frac{\varepsilon}{4}.$$

Because $\max\{a,b\} = \frac{1}{2}(a+b+|a-b|)$ it follows inductively from Step 1 that there exists $p \in \mathcal{P}(K \to \mathbb{R})$ with

$$\left\| \max_{1\le i\le N} p_{\delta,\mathbf{z}_i} - p \right\| < \varepsilon/4.$$

Using the triangular inequality for norms this implies

$$\|h - p\| < \varepsilon/2.$$

This completes Step 2 and the triangular inequality gives the overall desired result

$$\|f - p\| < \varepsilon,$$

which holds for any arbitrarily small $\varepsilon > 0$ and corresponding $p \in \mathcal{P}(K \to \mathbb{R})$. qed

B.3.2 Non-ε-denseness of $\mathcal{P}_N(K \to \mathbb{R})$ in $C(K \to \mathbb{R})$

It suffices to show the assertion for $K = [-1, 1]$, because any function in one variable can be viewed as a function in more than one variable, which does not depend on the other variables. If it is not possible to approximate a function $C([-1, 1] \to \mathbb{R})$ it will not be possible to approximate a function $f \in C(K \to \mathbb{R})$ for $K = [-1, 1]^{n_0}$ for $n_0 > 1$. Let $N \in \mathbb{N}$ and $\varepsilon > 0$, the aim is to construct a function $f \in C(K \to \mathbb{R})$, which can not be approximated with accuracy ε by any polynomial $p \in \mathcal{P}_N(K \to \mathbb{R})$ with degree N.

Let $x_0, x_1, \ldots, x_N \in [-1, 1]$ be pairwise distinct points. The function f, which will be constructed, will have the property $f(x_i) = 0$ for all $i \in \{0, \ldots, N\}$. To achieve an accuracy of ε the approximation polynomial must therefore fulfill $|p(x_i)| \leq \varepsilon$, i.e.

$$p \in \{\ q \in \mathcal{P}(K \to \mathbb{R}) \mid \exists\, y_0, y_1, \ldots, y_N \in [-\varepsilon, \varepsilon] : q(x_i) = y_i \text{ for all } i \in \{0, \ldots, N\}\ \}\,.$$

Clearly, for $\mathbf{y} := (y_0, y_1, \ldots, y_N) \in [-\varepsilon, \varepsilon]^{N+1}$,

$$q_{\mathbf{y}}(x) = \sum_{i=0}^{N} y_i \prod_{\substack{k=0 \\ k \neq i}}^{N} \frac{x - x_k}{x_i - x_k}$$

fulfills $q_{\mathbf{y}}(x_i) = y_i$ for all $i \in \{0, \ldots, N\}$ and it is furthermore the only polynomial of degree N with this property, because otherwise the difference between q and the other polynomial would be again a polynomial of N (or smaller) and it would have $N + 1$ zeros, which would imply by the Fundamental Theorem of Algebra (see, e.g., [Amann & Escher 2001*a*, Kor. I.8.18]) that the difference is the zero polynomial. From the definition of $q_{\mathbf{y}}$ it follows that the function

$$[-\varepsilon, \varepsilon]^{N+1} \times [-1, 1] \to \mathbb{R}, \quad (\mathbf{y}, x) \mapsto q_{\mathbf{y}}(x)$$

is continuous on the compact set $[-\varepsilon, \varepsilon]^N \times [-1, 1]$ and hence is bounded [Amann & Escher 2001*a*, Kor. 3.7.]. Let $B \in \mathbb{R}$ be such a bound, then

$$|p(x)| \leq B \quad \forall\, x \in [-1, 1]$$

for all polynomials $p \in \mathcal{P}_N(K \to \mathbb{R})$ which approximates f with accuracy ε. One possible definition of $f \in C(K \to \mathbb{R})$, which can not be approximated by any polynomial

with accuracy ε, is

$$f : [-1,1] \to \mathbb{R}, \quad x \mapsto \begin{cases} 0, & x \notin (x_0, x_1), \\ \frac{x-x_0}{x_1-x_0} 2(B+2\varepsilon) & x \in \left(x_0, \frac{x_0+x_1}{2}\right), \\ B+2\varepsilon & x = \frac{x_0+x_1}{2}, \\ \frac{x_1-x}{x_1-x_0} 2(B+2\varepsilon) & x \in \left(\frac{x_0+x_1}{2}, x_1\right), \end{cases}$$

which is continuous and for $x^* = \frac{x_0+x_1}{2}$ fulfills $f(x^*) = B + 2\varepsilon$. All polynomials p with $\big|p(x_i) - f(x_i)\big| \leq \varepsilon$ for all $i = 0, \ldots, N$, are bounded by B and therefore

$$\big|f(x^*) - p(x^*)\big| \geq 2\varepsilon > \varepsilon,$$

which implies the assertion. qed

B.4 Proof of Theorem 3.2.1

A detailed proof is given in [Pinkus 1999], therefore only the main steps are summarized here.

Let $K_1 = [-1,1]$ and $K = [-1,1]^{n_0}$.

Step 1: It is shown that $\mathcal{F}_{(1,(1,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$ *is dense in* $C(K_1 \to \mathbb{R})$ *for every* $\sigma \in C^\infty(K_1 \to \mathbb{R}) \backslash \mathcal{P}(K_1 \to \mathbb{R})$.
This step is Proposition 3.4 in [Pinkus 1999]. It is there shown that all polynomials are contained in the closure of $\mathcal{F}_{(1,(1,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$ and the claim then follows from the denseness of all polynomials in $C(K_1 \to \mathbb{R})$. This is done by first observing that if $\sigma \in C^\infty(K_1 \to \mathbb{R}) \backslash \mathcal{P}(K_1 \to \mathbb{R})$ then there exists a $\theta_0 \in K_1$ with

$$\sigma^{(k)}(-\theta_0) \neq 0 \quad \text{for all } k \in \mathbb{N}.$$

Afterwards, it is observed that for all $k \in \mathbb{N}$

$$t \mapsto D^k\big(\lambda \mapsto \sigma(\lambda t - \theta_0)\big)(0) = t \mapsto t^k \sigma^{(k)}(-\theta_0)$$

is an element of the closure of $\mathcal{F}_{(1,(1,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$, which implies that all monomials

and hence all polynomials are in the closure of $\mathcal{F}_{(1,(1,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$.

Step 2: It is shown that $\mathcal{F}_{(1,(n_0,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$ *is dense in* $C(K_1 \to \mathbb{R})$ *for every* $\sigma \in C(K_1 \to \mathbb{R}) \backslash \mathcal{P}(K_1 \to \mathbb{R})$.
This step is Proposition 3.7 in [Pinkus 1999]. Instead of σ it is

$$\sigma_\phi(t) = \int_{\mathbb{R}} \sigma(t-y)\phi(y)\mathrm{d}y$$

considered, where $\phi \in C^\infty(\mathbb{R} \to \mathbb{R})$ for which $\{\ y \in \mathbb{R} \mid \phi(y) \neq 0\ \}$ is bounded. It is $\sigma_\phi \in C^\infty$ and it is shown that the closure of $\mathcal{F}_{(1,(n_0,\cdot),\sigma_\phi,\cdot)}(K \to \mathbb{R})$ is contained in the closure of $\mathcal{F}_{(1,(n_0,\cdot),\sigma,\cdot)}(K \to \mathbb{R})$. Seeking a contradiction assume that there exists a $k \in \mathbb{N}$ for which $t \mapsto t^k$ is not an element of the closure of $\mathcal{F}_{(1,(n_0,\cdot),\sigma,\cdot)}(K \to \mathbb{R})$. This would imply that $\sigma_\phi^{(k)}(-\theta) = 0$ for all ϕ and all θ, i.e. σ_ϕ is a polynomial of degree at most $k-1$. Since there exists a sequence (ϕ_n) such that σ_{ϕ_n} converges uniformly to σ and the space of polynomials of degree at most $k-1$ is a closed subspace, σ would be a polynomial of degree at most $k-1$, too. This is a contradiction to the premise that σ is not a polynomial.

Step 3: It is shown that $\mathcal{F}_{(1,(n_0,\cdot),\sigma,\cdot)}(K \to \mathbb{R})$ *is dense in* $C(K \to \mathbb{R})$ *for every* $\sigma \in C(K \to \mathbb{R}) \backslash \mathcal{P}(K \to \mathbb{R})$.
This step is a specialized version of Proposition 3.3 in [Pinkus 1999]. To prove that the denseness in the one dimensional case carries over to the higher dimensional case, so called *ridge functions* are considered. These are function of the form

$$F(\mathbf{x}) = f(\mathbf{a} \cdot \mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{R}^{n_0}$$

where $f \in C(\mathbb{R} \to \mathbb{R})$ and $\mathbf{a} \in \mathbb{R}^{n_0}$, i.e. ridge functions are constant on the hyper planes $\mathbf{a} \cdot \mathbf{x} = c \in \mathbb{R}$. It can be shown that the set of all ridge functions is dense in $C(K \to \mathbb{R})$ (see, e.g., [Lin & Pinkus 1993]). Since ridge functions are actually one dimensional continuous functions they can itself be approximated by functions in $\mathcal{F}_{(1,(1,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$. The concatenation of a ridge function and a one dimensional function from $\mathcal{F}_{(1,(1,\cdot),\sigma,\cdot)}(K_1 \to \mathbb{R})$ is a function in $\mathcal{F}_{(n_0,(1,\cdot),\sigma,\cdot)}(K \to \mathbb{R})$ and therefore the denseness is shown. qed

B.5 Proof of Proposition 3.3.2

A detailed proof is given in [Pinkus 1999], therefore only the main aspects are summarized.

Step 1: Construction of activation function σ.

This step is based on Proposition 6.3 in [Pinkus 1999].

Let $\{\ u_k \in C^\infty([-1,1] \to \mathbb{R})\ \mid k \in \mathbb{N}_{>0}\ \}$ be a countable dense subset of $C([-1,1] \to \mathbb{R})$ (for example the set of all polynomials with rational coefficients). Let f be any strictly monotone differentiable function with $f(-\infty) = 0$ and $f(\infty) = 1$, for example the standard sigmoid function $t \mapsto \frac{1}{1+e^{-t}}$. For $m \in \mathbb{N}_{>0}$ define σ on the interval $[4m, 4m+2]$ by

$$\sigma(t + 4m + 1) = b_m + c_m t + d_m u_m(t) \quad \text{for } t \in [-1,1],$$

where $b_m, c_m, d_m \in \mathbb{R}$, $d_m \neq 0$, are chosen so that

(i) $\sigma(4m) = f(4m)$ and

(ii) $0 < \sigma'(t) \leq f'(t)$ for all $t \in [4m, 4m+2]$.

This is always possible since f is strictly increasing and $u_m{}'$ is bounded on $[-1,1]$ for all $m \in \mathbb{N}_{>0}$. On the intervals $[-4,-2]$ and $[0,2]$ chose σ to be linear and fulfilling conditions (i) and (ii) together with the condition that the function $t \mapsto \sigma(t-3)$ is linearly independent of the function $t \mapsto \sigma(t-1)$ on $[-1,1]$. Finally fill the gaps in the definition of σ such that σ is C^∞ and $\sigma(-\infty) = 0$. Since $\sigma(t + 4m + 1) = b_m + c_m t + d_m u_m(t)$ it follows that there exists $a_1^k, a_2^k, a_3^k \in \mathbb{R}$ with

$$a_1^k \sigma(t-3) + a_1^k \sigma(t-1) + a_3^k \sigma(t + 4m + 1) = u_m(t) \quad \text{for all } t \in [-1,1].$$

This property makes it possible for σ to approximate all continuous functions on $[-1,1]$ arbitrarily well.

Step 2: Showing that a finite number of hidden units suffices to approximate arbitrarily well.

This step is Theorem 7.1 in [Pinkus 1999]. Assume first that $K = [0,1]^{n_0}$, it is clear that the result then also holds for $K = [-1,1]^{n_0}$ by a simple coordinate transformation. It is a well known and famous result (Theorem of Kolmogorov, see, e.g., [Lorentz,

Golitschek & Makovoz 1996, Thm. 17.1.1]), that there exists $\lambda_1, \ldots, \lambda_{n_0} \in \mathbb{R}_{>0}$ with $\sum_{i=1}^{n_0} \lambda_i \leq 1$ and strictly increasing $\phi_1, \ldots, \phi_{2n_0+1} \in C([0,1] \to [0,1])$ such that for every continuous function $f \in C(K \to \mathbb{R})$ there exists a function $g \in C([0,1] \to \mathbb{R})$ such that

$$f(x_1, x_2, \ldots, x_{n_0}) = \sum_{i=1}^{2n_0+1} g\big(\lambda_1 \phi_i(x_1) + \lambda_2 \phi_i(x_2) + \ldots + \lambda_{n_0} \phi_i(x_{n_0})\big).$$

From Step 1 it follows that for every $\varepsilon > 0$ there exists $a_1, a_2, a_3 \in \mathbb{R}$ and $m \in \mathbb{N}$ such that

$$\Big| g(t) - \big(a_1 \sigma(t-3) + a_2 \sigma(t+1) + a_3 \sigma(t+m)\big) \Big| < \frac{\varepsilon}{2(2n_0+1)}$$

for all $t \in [0,1]$. Furthermore, for arbitrarily small $\delta > 0$ there exists $b_{i,1}, b_{i,2}, b_{i,3} \in \mathbb{R}$ and $r_i \in \mathbb{N}$ such that

$$\Big| \phi_i(x_j) - \big(b_{i,1} \sigma(x_j - 3) + b_{i,2} \sigma(x_j + 1) + b_{i,3} \sigma(x_j + r_i)\big) \Big| < \delta$$

for all $x_j \in [0,1]$. Choosing δ sufficiently small and substituting the equations into each others yield an approximation of f with an accuracy of ε by the MLP function of an MLP with two hidden layers and $4n+3$ units in the second layer and $2n+1$ units in the first hidden layer. qed

B.6 Theoretically possible approximation accuracy for certain function spaces

For $n_0 \in \mathbb{N}$ define

$$\overline{\mathbb{B}}_{n_0} := \left\{ \mathbf{x} \in \mathbb{R}^{n_0} \;\middle|\; \mathbf{x} = (x_1, x_2, \ldots, x_{n_0}) \text{ with } {x_1}^1 + {x_2}^2 + \ldots {x_{n_0}}^2 \leq 1 \right\},$$

i.e. $\overline{\mathbb{B}}_{n_0}$ is the compact unit ball with respect to the euclidian norm. Let the m-times continuously differentiable function space $C^m(\overline{\mathbb{B}}_{n_0} \to \mathbb{R})$ be equipped with the norm

$$\|f\|_{C^m} := \|f\| + \|f'\| + \ldots + \|f^{(m)}\|,$$

where all norms are the maximum norms on the corresponding function spaces (note that the values of f' and of the higher derivatives are not real numbers anymore,

see Subsection 4.1 for details). It is easy to see that $\|f\|_{C^m}$ is a norm, furthermore, $\left(C^m(\overline{\mathbb{B}}_{n_0} \to \mathbb{R}), \|\cdot\|_{C^m}\right)$ is a Banach space (see, e.g., [Werner 1995, Bsp. I.1.d]).

Proposition B.6.1. *Let σ be the standard sigmoid activation function, i.e. $\sigma(t) = \frac{1}{1+e^{-t}}$, and define for $n_0, n_1, m \in \mathbb{N}_{>0}$*

$$E(n_0, n_1, m) := \sup_{f \in C^m(\overline{\mathbb{B}}_{n_0} \to \mathbb{R})} \inf_{\substack{MLPs \\ (1,(n_0,n_1),\sigma,\cdot)}} \|f_{MLP} - f\|_{C^m},$$

i.e. $E(n_0, n_1, m)$ is the approximation capability of an MLP with n_1 hidden units in one layer for all functions in $C^m(\overline{\mathbb{B}}_{n_0} \to \mathbb{R})$. If $n_0 \geq 2$ then there exists constants $C_1, C_2 \in \mathbb{R}_{\geq 0}$, which are independent of n_1, such that

$$C_1(n_1 \log n_1)^{-m/n_0} \leq E(n_0, n_1, m) \leq C_2 {n_1}^{-m/n_0}.$$

This proposition is a specialization of Theorems 6.7 and 6.8 in [Pinkus 1999]. The main statement is that it is possible for single hidden layer MLP with a fixed number of hidden units to be ε-dense for arbitrarily small $\varepsilon > 0$, provided enough hidden units are used and the functions which should be approximated are at least differentiable. There seems to be no precise values for C_1 and C_2 and therefore these bounds can not be used to calculate how many hidden units are necessary. Nevertheless the bounds give a good orientation for the correlation between the number of hidden units and the theoretically possible approximation accuracy. It is for example possible to calculate how the accuracy increases if the number of hidden units is doubled.

B.7 Proof of Proposition 4.1.2

Step 1: Absolute convergence of $P(\mathbf{x})$
From (4.1.1) it follows inductively for $k \in \mathbb{N}$ that

$$\left|A_k(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x})\right| \leq \|A_k\| \, \|\mathbf{x}\|^k, \tag{B.7.1}$$

hence

$$\sum_{k=0}^{\infty} \left|A_k(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x})\right| \leq \sum_{k=0}^{\infty} \|A_k\| \, \|x\|^k \leq \sum_{k=0}^{\infty} \|A_k\|,$$

because $\|\mathbf{x}\| \leq 1$ for all $\mathbf{x} \in K$. Because $\limsup_{k\to\infty} \sqrt[k]{\|A_k\|} < 1$ the root criteria (see, e.g., [Amann & Escher 2001*a*, Thm. II.8.5]) yields

$$\sum_{k=0}^{\infty} \|A_k\| < \infty$$

and hence

$$\sum_{k=0}^{\infty} \big|A_k(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x})\big| < \infty.$$

Step 2: It is shown that $DP(\mathbf{x})$ exists and is given by

$$DP(\mathbf{x})\mathbf{h} = \sum_{k=1}^{\infty} kA_k(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{k-1 \text{ times}}, \mathbf{h}).$$

Step 2a: It is shown that $\sum_{k=1}^{\infty} kA_k(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{k-1 \text{ times}}, \mathbf{h})$ converges absolutely for all $\mathbf{x}, \mathbf{h} \in K$.

It is for all $\mathbf{x}, \mathbf{h} \in K$

$$\sum_{k=1}^{\infty} \big|kA_k(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{k-1 \text{ times}}, \mathbf{h})\big| \leq \sum_{k=1}^{\infty} k\|A_k\|$$

and

$$\limsup_{k\to\infty} \sqrt[k]{k\|A_k\|} = \lim \underbrace{\sqrt[k]{k}}_{=1} \limsup_{k\to\infty} \sqrt[k]{\|A_k\|} < 1,$$

hence the root criteria yields that $\sum_{k=1}^{\infty} kA_k(\mathbf{x}, \ldots, \mathbf{x}, \mathbf{h})$ converges absolutely.

Step 2b: It is shown that for all $k \in \mathbb{N}$

$$A_k(\mathbf{x}+\mathbf{h}, \mathbf{x}+\mathbf{h}, \ldots, \mathbf{x}+\mathbf{h}) = \sum_{i=0}^{k} \binom{k}{i} A(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{k \text{ times}}, \underbrace{\mathbf{h}, \ldots, \mathbf{h}}_{n-k \text{ times}}).$$

The proof is done inductively. For $k = 0$ the claim is obviously fulfilled. From the multilinearity of A_k it follows that

$$\begin{aligned} A_k(\mathbf{x}+\mathbf{h}, \mathbf{x}+\mathbf{h}, \ldots, \mathbf{x}+\mathbf{h}) = {} & A_k(\mathbf{x}+\mathbf{h}, \mathbf{x}+\mathbf{h}, \ldots, \mathbf{x}+\mathbf{h}, \mathbf{x}) \\ & + A_k(\mathbf{x}+\mathbf{h}, \mathbf{x}+\mathbf{h}, \ldots, \mathbf{x}+\mathbf{h}, \mathbf{h}), \end{aligned}$$

Fixing the last entry of the both right hand side terms the induction hypothesis yields

$$A_k(\mathbf{x}+\mathbf{h},\mathbf{x}+\mathbf{h},\dots,\mathbf{x}+\mathbf{h},\mathbf{a}) = \sum_{i=0}^{k-1}\binom{k-1}{i} A_k(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{k \text{ times}}, \underbrace{\mathbf{h},\dots,\mathbf{h}}_{n-1-k \text{ times}},\mathbf{a}),$$

where either $\mathbf{a}=\mathbf{x}$ or $\mathbf{a}=\mathbf{h}$. The symmetry of A_k yields

$$\begin{aligned} A_k(\mathbf{x}+\mathbf{h},\mathbf{x}+\mathbf{h},\dots,\mathbf{x}+\mathbf{h}) &= A_k(\mathbf{h},\mathbf{h},\dots,\mathbf{h}) \\ &+ \sum_{i=1}^{k-1}\left(\binom{k-1}{i-1}+\binom{k-1}{i}\right) A_k(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{i \text{ times}},\underbrace{\mathbf{h},\dots,\mathbf{h}}_{k-i \text{ times}}) + A_k(\mathbf{x},\mathbf{x},\dots,\mathbf{x}) \end{aligned}$$

and since $\binom{k-1}{i-1}+\binom{k-1}{i}=\binom{k}{i}$ Step 2b is shown.

Step 2c: The formula for $DP(\mathbf{x})$ is shown.
From Step 2b it follows that

$$P(\mathbf{x}+\mathbf{h}) = \sum_{k=0}^{\infty}\sum_{i=0}^{k}\binom{k}{i} A_k(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{i \text{ times}},\underbrace{\mathbf{h},\dots,\mathbf{h}}_{k-i \text{ times}})$$

and hence

$$P(\mathbf{x}+\mathbf{h}) - P(\mathbf{x}) = \sum_{k=1}^{\infty} kA_k(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{k-1 \text{ times}},\mathbf{h}) + \sum_{k=2}^{\infty}\sum_{i=0}^{k-2}\binom{k}{i} A_k(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{i \text{ times}},\underbrace{\mathbf{h},\dots,\mathbf{h}}_{k-i \text{ times}}).$$

Therefore, since $\|\mathbf{x}\|\le 1$,

$$\begin{aligned} \left|\frac{P(\mathbf{x}+\mathbf{h})-P(\mathbf{x})-\sum_{k=1}^{\infty}kA_k(\mathbf{x},\dots,\mathbf{x},\mathbf{h})}{\|\mathbf{h}\|}\right| &\le \|\mathbf{h}\|\sum_{k=2}^{\infty}\sum_{i=0}^{k-2}\binom{k}{i}\|A_k\|\,\|\mathbf{h}\|^{k-2-i} \\ &\le \|\mathbf{h}\|\sum_{k=2}^{\infty}\|A_k\|k(k-1)(1+\|\mathbf{h}\|)^k. \end{aligned}$$

Let

$$\rho := \limsup_{k\to\infty}\sqrt[k]{\|A_k\|k(k-1)} = \limsup_{k\to\infty}\sqrt[k]{\|A_k\|}\underbrace{\lim_{k\to\infty}\sqrt[k]{k(k-1)}}_{=1} < 1$$

and chose $\varepsilon > 0$ such that $\rho(1+\varepsilon) < 1$, then

$$\limsup_{k\to\infty} \sqrt[k]{\|A_k\|k(k-1)(1+\varepsilon)^k} < 1,$$

and hence

$$C_\varepsilon := \sum_{k=2}^{\infty} \|A_k\|k(k-1)(1+\varepsilon)^k < \infty.$$

Finally,

$$0 \le \lim_{\mathbf{h}\to 0} \left| \frac{P(\mathbf{x}+\mathbf{h}) - P(\mathbf{x}) - \sum_{k=1}^{\infty} kA_k(\mathbf{x},\dots,\mathbf{x},\mathbf{h})}{\|\mathbf{h}\|} \right| \le \lim_{\mathbf{h}\to 0} \|\mathbf{h}\| C_\varepsilon = 0,$$

which yields $DP(x) = \sum_{k=1}^{\infty} kA_k(\mathbf{x},\dots,\mathbf{x},\mathbf{h})$, hence Step 2 is finished.

Step 3: It is shown that $D^n P(0) = n! A_n$.

Step 3a: It is shown that

$$D^n P(\mathbf{x})(\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_n) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1)A_k(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{k-n \text{ times}},\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_n).$$

This step is proved inductively. The case $n = 1$ was already shown in Step 2. It is for fixed $\mathbf{h}_1,\dots,\mathbf{h}_{n+1} \in K$

$$D^{n+1}P(\mathbf{x})(\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_{n+1}) = D\big(D^n P(\mathbf{x})(\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_n)\big)(\mathbf{h}_{n+1})$$

and $P_n(\mathbf{x}) := D^n P(\mathbf{x})(\mathbf{h}_1,\dots,\mathbf{h}_n)$ can be written as

$$P_n(\mathbf{x}) = \sum_{i=0}^{\infty} B_i(\mathbf{x},\dots,\mathbf{x}),$$

where $B_i(\mathbf{x},\dots,\mathbf{x}) = (i+n)(i+n-1)\cdots(i+1)A_{n+i}(\mathbf{x},\dots,\mathbf{x},\mathbf{h}_1,\dots,\mathbf{h}_n)$. Invoking Step 2 again yields

$$DP_n(\mathbf{x})(\mathbf{h}_{n+1}) = \sum_{i=1}^{\infty} iB_i(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{i-1 \text{ times}},\mathbf{h}_{n+1}),$$

hence by the symmetry of A_k for $k \in \mathbb{N}$

$$D^{n+1}P(\mathbf{x})(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n+1}) = \sum_{k=n+1}^{\infty} (k-n) B_{k-n}(\mathbf{x}, \ldots, \mathbf{x}, \mathbf{h}_{n+1})$$
$$= \sum_{k=n+1}^{\infty} k(k-1)\cdots(k-n+1)(k-n) A_k(\mathbf{x}, \ldots, \mathbf{x}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n, \mathbf{h}_{n+1}).$$

Step 3b: The formula for $D^nP(0)$ is shown.
It is by Step 3a

$$D^nP(0)(\mathbf{h}_1, \ldots, \mathbf{h}_n)$$
$$= n! A_n(\mathbf{h}_1, \ldots, \mathbf{h}_n) + \sum_{k=n+1}^{\infty} k(k-1)\cdots(k-n+1) A_k(0, \ldots, 0, \mathbf{h}_1, \ldots, \mathbf{h}_n).$$

Because A_k is multilinear it is $A_k(0, \ldots, 0, \mathbf{h}_1, \ldots, \mathbf{h}_n) = 0$ for all $k \geq n+1$. Hence Step 3 is shown and Proposition 4.1.2 is proven. qed

B.8 Proof of Lemma 4.5.3

Let $F : K \to \mathbb{R}$ be given by

$$F(\mathbf{x}) = w_1 f_1(\mathbf{x}) + \ldots + w_m f_m(\mathbf{x}) + \theta.$$

Since f_i are nicely analytical they can be expressed through their Taylor series and hence

$$F(\mathbf{x}) = \sum_{k=0}^{\infty} \left(\theta_k + \sum_{i=1}^{m} w_i \frac{D^k f_i(0)}{k!} \right) \mathbf{x}^k,$$

where $\theta_0 \mathbf{x}^0 := \theta$ and $\theta_k \mathbf{x}^k = 0$ for all $k \geq 1$, i.e. the Taylor series of F converges to F on the whole of K. The function F is also nicely analytical, because

$$\begin{aligned}
\limsup_{k\to\infty} \sqrt[k]{\frac{\|D^k F(0)\|}{k!}} &= \limsup_{k\to\infty} \sqrt[k]{\frac{\|\theta_k + \sum_{i=1}^m w_i D^k f_i(0)\|}{k!}} \\
&\le \limsup_{k\to\infty} \sqrt[k]{\frac{\sum_{i=1}^m |w_i| \big\|D^k f_i(0)\big\|}{k!}} \\
&\le \max_{i=1..m} \limsup_{k\to\infty} \sqrt[k]{m|w_i| \frac{\big\|D^k f_i(0)\big\|}{k!}} \\
&= \max_{i=1..m} \underbrace{\lim_{k\to\infty} \sqrt[k]{m|w_i|}}_{=1} \underbrace{\limsup_{k\to\infty} \sqrt[k]{\frac{\big\|D^k f_i(0)\big\|}{k!}}}_{<1} \\
&< 1.
\end{aligned}$$

It must be shown now that $g(\mathbf{x}) = \sigma\big(F(\mathbf{x})\big)$ can be written as

$$g(\mathbf{x}) = \sum_{k=0}^{k} A_k(\mathbf{x}, \dots, \mathbf{x})$$

for some $A_k \in \mathcal{L}^k_{\text{sym}}(K \to \mathbb{R})$, $k \in \mathbb{N}$ with $\limsup_{k\to\infty} \sqrt[k]{\|A_k\|} < 1$, because Proposition 4.1.2 would then yield that g is nicely analytical.
The following proof is similar to the proof in [Walter 2004, Satz 7.13], where classical power series are considered. It is

$$g(\mathbf{x}) = \sum_{k=0}^{k} a_k \left(\sum_{i=0}^{\infty} \frac{D^i F(0) \mathbf{x}^i}{i!} \right)^k .$$

Claim: For every $k \in \mathbb{N}$ there exists $B_i^k \in \mathcal{L}^k_{sym}(K \to \mathbb{R})$, $i \in \mathbb{N}$ with

$$\left(\sum_{i=0}^{\infty} \frac{D^i F(0) \mathbf{x}^i}{i!} \right)^k = \sum_{i=0}^{\infty} B_i^k (\underbrace{\mathbf{x}, \dots, \mathbf{x}}_{i \text{ times}})$$

and the right hand side series converges absolutely for all $\mathbf{x} \in K$.

The claim is proved by induction. For $k = 1$ nothing is to show. Now

$$\left(\sum_{i=0}^{\infty} \frac{D^i F(0) \mathbf{x}^i}{i!} \right)^{k+1} = \sum_{i=0}^{\infty} B_i^k(\mathbf{x}, \dots, \mathbf{x}) \left(\sum_{i=0}^{\infty} \frac{D^i F(0) \mathbf{x}^i}{i!} \right),$$

where for fixed $\mathbf{x} \in K$ both series converge absolutely and hence the Cauchy product

$$\sum_{i=0}^{\infty} \sum_{n=0}^{i} B_n^k(\mathbf{x}, \dots, \mathbf{x}) \frac{D^{i-n}F(0)\mathbf{x}^{i-n}}{(i-n)!}$$

converges also absolutely (see, e.g., [Amann & Escher 2001*a*, Thm. II.8.11]). Let for $i \in \mathbb{N}$

$$\tilde{B}_i^{k+1}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i) := \sum_{n=0}^{i} B_n^k(\mathbf{x}_1, \dots, \mathbf{x}_n) \frac{D^{i-n}F(0)(\mathbf{x}_{n+1}, \dots, \mathbf{x}_i)}{(i-n)!}$$

then $\tilde{B}_i^{k+1}$ is multilinear. In general $\tilde{B}_i^{k+1}$ will not be symmetric, but it is possible to construct a symmetric and multilinear B_i^{k+1} from $\tilde{B}_i^{k+1}$ with $B_i^{k+1}(\mathbf{x}, \dots, \mathbf{x}) = \tilde{B}_i^{k+1}(\mathbf{x}, \dots, \mathbf{x})$, see [Dineen 1999, p. 6]. Therefore the claim is shown.

Let

$$F_{\text{abs}}(x) := \sum_{i=0}^{\infty} \frac{\left\|D^i F(0)\right\|}{i!} x^i,$$

then, compare (B.7.1), the absolute series of $F(\mathbf{x})$ is bounded by $F_{\text{abs}}\big(\|\mathbf{x}\|\big)$. By [Amann & Escher 2001*b*, Thm. II.9.2] the (classical) power series F_{abs} converges absolutely for all $x \in \mathbb{R}$ with

$$|x| < \delta := \underbrace{1/\limsup_{k\to\infty} \sqrt[i]{\frac{\left\|D^i F(0)\right\|}{i!}}}_{>1}.$$

With the same inductive proof as above it can now be shown that for every $k \in \mathbb{N}$

$$F_{\text{abs}}(x)^k = \sum_{i=0}^{\infty} b_i^k x^i,$$

for some $b_i^k \in \mathbb{R}_{\geq 0}$, $i \in \mathbb{N}$, in particular

$$\|B_i^k\| \leq b_i^k,$$

because the values for the b_i^k are calculated from $\frac{\left\|D^i F(0)\right\|}{i!}$ in the same way as the B_i^k are calculated from $\frac{D^i F(0)}{i!}$. Let

$$g_{\text{abs}}(x) := \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} |a_k| b_i^k x^k = \sum_{k=0}^{\infty} |a_k| F_{\text{abs}}(x)^k,$$

then $g_{\text{abs}}(x)$ converges for every $x \in \mathbb{R}$ with $|x| < \rho$, because $\sigma(t) = \sum_{k=0}^{\infty} a_k t^k$ converges absolutely for all $t \in \mathbb{R}$. Therefore, for all $\mathbf{x} \in K$ (i.e. $\|\mathbf{x}\| \leq 1 < \rho$),

$$\sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \left| a_k B_i^k(\mathbf{x}, \ldots, \mathbf{x}) \right| \leq \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} |a_k| \, b_i^k \, \|\mathbf{x}\|^i = g_{\text{abs}}\big(\|\mathbf{x}\|\big) < \infty,$$

hence, the summation order in

$$g(\mathbf{x}) = \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} a_k B_i^k(\mathbf{x}, \ldots, \mathbf{x})$$

can be changed, see, e.g., [Amann & Escher 2001*a*, Thm. II.8.10]:

$$g(\mathbf{x}) = \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} a_k B_i^k(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{i \text{ times}}),$$

which shows that

$$g(\mathbf{x}) = \sum_{i=0}^{k} A_i(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x}),$$

where $A_i \in \mathcal{L}^i_{\text{sym}}(K \to \mathbb{R})$ is given by

$$A_i(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i) := \sum_{k=0}^{\infty} a_k B_i^k(\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_i).$$

It remains to show that $\limsup_{k\to\infty} \sqrt[k]{\|A_k\|} < 1$.

The radius of convergence of the power series g_{abs} is bigger than one, and therefore, by [Amann & Escher 2001*b*, Thm. II.8.5],

$$\limsup_{i\to\infty} \sqrt[i]{\sum_{k=0}^{\infty} |a_k| b_i^k} < 1.$$

From

$$\|A_i\| \leq \sum_{k=0}^{\infty} |a_k| b_i^k$$

it follows that

$$\limsup_{i\to\infty} \sqrt[i]{\|A_i\|} < 1$$

and the proof of Lemma 4.5.3 is complete. qed

B.9 Proof of Lemma 4.5.4 and table of derivatives of the sigmoid activation function

The proof of Lemma 4.5.4 is done by induction. It is easy to see that the assertion of Lemma 4.5.4 is true for $n = 1$. Now, for all $n \geq 1$,

$$\begin{aligned}
\sigma^{(n+1)} &= \left(\sigma^{(n)}\right)' = \sum_{i=1}^{n+1} a_{i,n}\,(\sigma^i)' \\
&= \sum_{i=1}^{n+1} i\,a_{i,n}\,\sigma^{i-1}\,\sigma' \\
&= \sum_{i=1}^{n+1} i\,a_{i,n}\,\sigma^{i-1}\,(\sigma-\sigma^2) \\
&= a_{1,n}\sigma + \sum_{i=2}^{n+1} (i\,a_{i,n} - (i-1)\,a_{i-1,n})\,\sigma^i + (n+1)\,a_{n+1,n}\sigma^{n+2}.
\end{aligned}$$

It must be shown that for all $n \in \mathbb{N}_{>0}$

(i) $a_{1,n+1} = a_{1,n}$,

(ii) $a_{i,n+1} = i\,a_{i,n} - (i-1)\,a_{i-1,n}$, for all $i = 2, \ldots, n+1$, and

(iii) $a_{n+2,n+1} = -(n+1)\,a_{n+1,n}$.

It is $a_{1,n+1} = 1 = a_{1,n}$ and

$$\begin{aligned}
a_{i,n}i - a_{i-1,n}(i-1) &= \sum_{k=1}^{i} (-1)^{k+1} i \binom{i-1}{k-1} k^n - \sum_{k=1}^{i-1} (-1)^{k+1} (i-1) \binom{i-2}{k-1} k^n \\
&= \sum_{k=1}^{i-1} (-1)^{k+1} k^{n+1} \left(\frac{i}{k} \binom{i-1}{k-1} - \frac{i-1}{k} \binom{i-2}{k-1} \right) \\
&\quad + (-1)^{i+1} \binom{i-1}{i-1} i^{n+1}
\end{aligned}$$

for all $n \in \mathbb{N}$ and $i \in \{2, \ldots, n+1\}$. From

$$\frac{i}{k}\binom{i-1}{k-1} - \frac{i-1}{k}\binom{i-2}{k-1} = \frac{i!}{(i-k)!k!} - \frac{(i-1)!}{(i-k-1)!k!} = \frac{i! - (i-k)(i-1)!}{(i-k)!k!}$$
$$= \frac{(i-1)!}{(i-k)!(k-1)!} = \binom{i-1}{k-1},$$

it follows that

$$a_{i,n}i - a_{i-1,n}(i-1) = \sum_{k=1}^{i}(-1)^{k+1}\binom{i-1}{k-1}k^{n+1} = a_{i,n+1},$$

and only (iii) remains to be shown.
Because $a_{2,1} = -1$, showing (iii) is equivalent to showing that

$$a_{n+1,n} = (-1)^n n!,$$

i.e.

$$n! = (-1)^n \sum_{k=1}^{n+1}(-1)^{(k+1)}\binom{n}{k-1}k^n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}(-1-k)^n.$$

The latter follows from Ruiz identity [Ruiz 1996]:

$$n! = \sum_{k=0}^{n}(-1)^k\binom{n}{k}(x-k)^n, \quad \forall\, x \in \mathbb{R}.$$

qed

$\mathbf{n}$	$\sigma^{(\mathbf{n})}$	$\sigma^{(\mathbf{n})}(\mathbf{0})$
1	$\sigma - \sigma^2$	1/4
2	$\sigma - 3\,\sigma^2 + 2\,\sigma^3$	0
3	$\sigma - 7\,\sigma^2 + 12\,\sigma^3 - 6\,\sigma^4$	-1/8
4	$\sigma - 15\,\sigma^2 + 50\,\sigma^3 - 60\,\sigma^4 + 24\,\sigma^5$	0
5	$\sigma - 31\,\sigma^2 + 180\,\sigma^3 - 390\,\sigma^4 + 360\,\sigma^5 - 120\,\sigma^6$	1/4
6	$\sigma - 63\,\sigma^2 + 602\,\sigma^3 - 2100\,\sigma^4 + 3360\,\sigma^5 - 2520\,\sigma^6 + 720\,\sigma^7$	0
7	$\sigma - 127\,\sigma^2 + 1932\,\sigma^3 - 10206\,\sigma^4 + 25200\,\sigma^5 - 31920\,\sigma^6 + 20160\,\sigma^7 - 5040\,\sigma^8$	-17/16
8	$\sigma - 255\,\sigma^2 + 6050\,\sigma^3 - 46620\,\sigma^4 + 166824\,\sigma^5 - 317520\,\sigma^6 + 332640\,\sigma^7 - 181440\,\sigma^8 + 40320\,\sigma^9$	0
9	$\sigma - 511\,\sigma^2 + 18660\,\sigma^3 - 204630\,\sigma^4 + 1020600\,\sigma^5 - 2739240\,\sigma^6 + 4233600\,\sigma^7 - 3780000\,\sigma^8 + 1814400\,\sigma^9 - 362880\,\sigma^{10}$	31/4
10	$\sigma - 1023\,\sigma^2 + 57002\,\sigma^3 - 874500\,\sigma^4 + 5921520\,\sigma^5 - 21538440\,\sigma^6 + 46070640\,\sigma^7 - 59875200\,\sigma^8 + 46569600\,\sigma^9 - 19958400\,\sigma^{10} + 3628800\,\sigma^{11}$	0
11	$\sigma - 2047\,\sigma^2 + 173052\,\sigma^3 - 3669006\,\sigma^4 + 33105600\,\sigma^5 - 158838240\,\sigma^6 + 451725120\,\sigma^7 - 801496080\,\sigma^8 + 898128000\,\sigma^9 - 618710400\,\sigma^{10} + 239500800\,\sigma^{11} - 39916800\,\sigma^{12}$	-691/8
12	$\sigma - 4095\,\sigma^2 + 523250\,\sigma^3 - 15195180\,\sigma^4 + 180204024\,\sigma^5 - 1118557440\,\sigma^6 + 4115105280\,\sigma^7 - 9574044480\,\sigma^8 + 14495120640\,\sigma^9 - 14270256000\,\sigma^{10} + 8821612800\,\sigma^{11} - 3113510400\,\sigma^{12} + 479001600\,\sigma^{13}$	0
13	$\sigma - 8191\,\sigma^2 + 1577940\,\sigma^3 - 62350470\,\sigma^4 + 961800840\,\sigma^5 - 7612364760\,\sigma^6 + 35517081600\,\sigma^7 - 105398092800\,\sigma^8 + 207048441600\,\sigma^9 - 273158645760\,\sigma^{10} + 239740300800\,\sigma^{11} - 134399865600\,\sigma^{12} + 43589145600\,\sigma^{13} - 6227020800\,\sigma^{14}$	5461/4
14	$\sigma - 16383\,\sigma^2 + 4750202\,\sigma^3 - 254135700\,\sigma^4 + 5058406080\,\sigma^5 - 50483192760\,\sigma^6 + 294293759760\,\sigma^7 - 1091804313600\,\sigma^8 + 2706620716800\,\sigma^9 - 4595022432000\,\sigma^{10} + 5368729766400\,\sigma^{11} - 4249941696000\,\sigma^{12} + 2179457280000\,\sigma^{13} - 653837184000\,\sigma^{14} + 87178291200\,\sigma^{15}$	0
15	$\sigma - 32767\,\sigma^2 + 14283372\,\sigma^3 - 1030793406\,\sigma^4 + 26308573200\,\sigma^5 - 328191186960\,\sigma^6 + 2362955474880\,\sigma^7 - 10794490827120\,\sigma^8 + 33094020960000\,\sigma^9 - 70309810771200\,\sigma^{10} + 105006251750400\,\sigma^{11} - 110055327782400\,\sigma^{12} + 79332244992000\,\sigma^{13} - 37486665216000\,\sigma^{14} + 10461394944000\,\sigma^{15} - 1307674368000\,\sigma^{16}$	-929569/32

Table 4: Derivates of the sigmoid activation function given by $\sigma(t) = \frac{1}{1+e^{-t}}$.

References

Abraham, R., Marsden, J. & Ratiu, T. [1988], *Manifolds, Tensor Analysis, and Applications*, number 75 *in* 'Applied Mathematical Sciences', 2nd edn, Springer-Verlag, New York.

Amann, H. & Escher, J. [2001*a*], *Analysis I*, Birkhäuser Verlag, Basel - Boston - Berlin.

Amann, H. & Escher, J. [2001*b*], *Analysis II*, Birkhäuser Verlag, Basel - Boston - Berlin.

Barron, A. [1994], 'Approximation and estimation bounds for artificial neural networks', *Machine Learning* **14**, 115–133.

Dineen, S. [1999], *Complex analysis on infinite dimensional spaces*, Springer monographs in mathematics, Springer-Verlag, London.

Haykin, S. [1994], *Neural Networks*, Macmillan College Publishing Company, New York.

Lin, V. & Pinkus, A. [1993], 'Fundamentality of ridge functions', *Journal of Approximation Theory* **75**(3), 295–311.

Lorentz, G., Golitschek, M. & Makovoz, Y. [1996], *Constructive approximation: advanced problems*, number 304 *in* 'Grundlehren der mathematischen Wissenschaften', Springer-Verlag, Berlin Heidelberg.

Merker, J. [2006], Private communication.

Nørgaard, M., Ravn, O., Poulsen, N. & Hansen, L. [2000], *Neural Networks for Modelling and Control of Dynamic Systems*, Advanced Textbooks in Control and Signal Processing, Springer-Verlag, London.

Pinkus, A. [1999], 'Approximation theory of the MLP model in neural networks', *Acta Numerica* **9**, 143–195.

Pinkus, A. [2006], Private communication.

Ruiz, S. [1996], 'An algebraic identity leading to Wilsons theorem', *The Mathematical Gazette* **80**(489), 579–582.

Rumelhart, D., Hinton, G. & Williams, R. [1986], Learning internal representation by error propagation, *in* J. Feldman, P. Hayes & D. Rumelhart, eds, 'Parallel Distributed Processing: Explorations in the Microstructure of Cognition', Vol. 1 of *Computational Models of Cognition and Perception*, The MIT Press, Cambridge London, chapter 8, pp. 318–362.

Sagan, H. [1994], *Space-filling curves*, Springer-Verlag, New York.

Sarle, W. [2002], 'Neural networks, FAQ, parts 1 to 7', Usenet newsgroup `comp.ai.neural-nets`, `ftp://ftp.sas.com/pub/neural/FAQ.html`.

Walter, W. [2004], *Analysis 1*, 7th edn, Springer-Verlag, Berlin Heidelberg New York.

Werner, D. [1995], *Funktionalanalysis*, Springer-Verlag, Berlin Heidelberg.