

Mathematical Introduction to Deep Neural Networks

Lecture Notes
Diyora Salimova
University of Freiburg

Preface

These lecture notes are slightly modified version of an excerpt from the book [4]. The authors of [4] are gratefully acknowledged for agreeing to the use of [4] in making current lecture notes.

These lecture notes aims to provide an introduction to the topic of deep learning algorithms. Roughly speaking, when we speak of a *deep learning algorithm* we think of a computational scheme which aims to approximate certain relations, functions, or quantities by means of so-called deep *artificial neural networks* (ANNs) and the iterated use of some kind of data. ANNs, in turn, can be thought of as classes of functions that consist of multiple compositions of certain nonlinear functions, which are referred to as *activation functions*, and certain affine functions. Loosely speaking, the depth of such ANNs corresponds to the number of involved iterated compositions in the ANN and one starts to speak of *deep* ANNs when the number of involved compositions of nonlinear and affine functions is larger than two.

After a brief [introduction](#), these lecture notes are divided into three parts (see Parts [I](#), [II](#), and [III](#)). In Part [I](#) we introduce in Chapter [1](#) *fully-connected feedforward ANNs*, in all mathematical details and in Chapter [2](#) we present a certain calculus for fully-connected feedforward ANNs. For the sake of conciseness, we use the term ANN when we actually mean fully-connected feedforward ANN.

In Part [II](#) we present several mathematical results that analyze how well ANNs can approximate given functions. To make this part more accessible, we restrict ourselves to one-dimensional functions from the reals to the reals.

A key aspect of deep learning algorithms is usually to model or reformulate the problem under consideration as a suitable optimization problem involving deep ANNs. It is precisely the subject of Part [III](#) to study such and related optimization problems and the corresponding optimization algorithms to approximately solve such problems in detail. In particular, in the context of deep learning methods such optimization problems – typically given in the form of a minimization problem – are usually solved by means of appropriate *gradient based* optimization methods. Roughly speaking, we think of a gradient based optimization method as a computational scheme which aims to solve the considered optimization problem by performing successive steps based on the direction of the (negative) gradient of the function which one wants to optimize. Deterministic variants of such gradient based optimization methods such as the *gradient descent* (GD) optimization method are reviewed and studied in Chapter [5](#) and stochastic variants of such gradient based optimization methods such as the *stochastic gradient descent* (SGD) optimization method are reviewed and studied in Chapter [6](#). GD-type and SGD-type optimization methods can, roughly speaking, be viewed as time-discrete approximations of solutions of suitable *gradient flow* (GF) *ordinary differential equations* (ODEs). To develop intuitions for GD-type and SGD-type optimization methods and for some of the tools which we employ to analyze such methods, we study in Chapter [4](#) such GF ODEs. In particular, we show in Chapter [4](#) how such GF ODEs can be

used to approximately solve appropriate optimization problems.

The mathematical analyses for gradient based optimization methods that we present in Chapters 4, 5, and 6 are in almost all cases too restrictive to cover optimization problems associated to the training of ANNs. However, such optimization problems can be covered by the so called *Kurdyka–Łojasiewicz* (KL) approach. Popular methods that aim to accelerate ANN training procedures in data-driven learning problems are so called *batch normalization* (BN) methods which are not subject of these lecture notes.

The mathematical analysis of deep learning algorithms does not only consist of error estimates for approximation capacities of ANNs (cf. Part II) and of error estimates for the involved optimization methods (cf. Part III) but also requires estimates for the *generalization error* which, roughly speaking, arises when the probability distribution associated to the learning problem cannot be accessed explicitly but is approximated by a finite number of realizations/data. The analysis of this error is not subject of these lecture notes.

One then can combine the *approximation error* estimates, the *optimization error* estimates, and the *generalization error* estimates to establish estimates for the overall error in the exemplary situation of the training of ANNs based on SGD-type optimization methods with many independent random initializations.

Contents

Preface	2
Introduction	7
I Artificial neural networks (ANNs)	11
1 Basics on ANNs	12
1.1 Fully-connected feedforward ANNs (vectorized description)	12
1.1.1 Affine functions	14
1.1.2 Vectorized description of ANNs	15
1.1.3 Weight and bias parameters of ANNs	17
1.2 Activation functions	19
1.2.1 Multidimensional versions	19
1.2.2 Single hidden layer ANNs	20
1.2.3 Rectified linear unit (ReLU) activation	21
1.2.4 Other activation functions	24
1.3 ANNs (structured description)	29
1.3.1 Structured description of ANNs	29
1.3.2 Realizations of ANNs	31
1.3.3 On the connection to the vectorized description	32
2 ANN calculus	36
2.1 Compositions of ANNs	36
2.1.1 Compositions of ANNs	36
2.1.2 Elementary properties of compositions of ANNs	37
2.1.3 Associativity of compositions of ANNs	38
2.1.4 Powers of ANNs	39
2.2 Parallelizations of ANNs	39
2.2.1 Parallelizations of ANNs with the same length	39
2.2.2 Representations of the identities with ReLU activation functions	44
2.2.3 Extensions of ANNs	45

2.2.4	Parallelizations of ANNs with different lengths	49
2.3	Scalar multiplications of ANNs	51
2.3.1	Affine transformations as ANNs	51
2.3.2	Scalar multiplications of ANNs	53
2.4	Sums of ANNs with the same length	54
2.4.1	Sums of vectors as ANNs	54
2.4.2	Concatenation of vectors as ANNs	56
2.4.3	Sums of ANNs	58
II Approximation		61
3 One-dimensional ANN approximation results		62
3.1	Linear interpolation of one-dimensional functions	62
3.1.1	On the modulus of continuity	62
3.1.2	Linear interpolation of one-dimensional functions	64
3.2	Linear interpolation with ANNs	68
3.2.1	Activation functions as ANNs	68
3.2.2	Representations for ReLU ANNs with one hidden neuron	70
3.2.3	ReLU ANN representations for linear interpolations	71
3.3	ANN approximations results for one-dimensional functions	74
3.3.1	Constructive ANN approximation results	74
3.3.2	Convergence rates for the approximation error	78
III Optimization		82
4 Optimization through gradient flow (GF) trajectories		83
4.1	Introductory comments for the training of ANNs	83
4.2	Loss functions	87
4.2.1	Absolute error loss	87
4.2.2	Mean squared error loss	88
4.3	GF optimization in the training of ANNs	89
4.4	Lyapunov-type functions for GFs	89
4.4.1	Gronwall differential inequalities	89
4.4.2	Lyapunov-type functions for ODEs	91
4.4.3	Sufficient and necessary conditions for local minimum points	92
4.4.4	On a linear growth condition	95
4.5	Optimization through flows of ODEs	95
4.5.1	Approximation of local minimum points through GFs	95

5	Deterministic gradient descent (GD) optimization methods	99
5.1	GD optimization	99
5.1.1	GD optimization in the training of ANNs	100
5.1.2	Euler discretizations for GF ODEs	101
5.1.3	Lyapunov-type stability for GD optimization	102
5.1.4	Error analysis for GD optimization	105
6	Stochastic gradient descent (SGD) optimization methods	115
6.1	Introductory comments for the training of ANNs with SGD	115
6.2	SGD optimization	117
6.2.1	SGD optimization in the training of ANNs	118
6.2.2	Convergence rates for SGD for coercive objective functions	120
7	List of definitions	123
	Bibliography	125

Introduction

Roughly speaking, the field *deep learning* can be divided into three subfields, deep *supervised learning*, deep *unsupervised learning*, and deep *reinforcement learning*. Each of these approaches corresponds to a different type of available data and feedback.

Supervised learning deals with *labeled data* — data for which both the input and the desired output are known. Examples can be predicting house prices and classifying handwritten digits. Supervised learning can be described as *learning by example*: the algorithm learns from input–output pairs how to generalize to new data.

Unsupervised learning works with *unlabeled data* — only inputs are provided, without corresponding outputs. The goal is to uncover hidden structure or patterns within the data. Examples are customer segmentation (group customers into segments based on purchasing behavior, demographics, or browsing history) and image segmentation (partition images into meaningful regions). Unsupervised learning can be viewed as *learning by observation*: the algorithm explores the data’s internal structure without any explicit feedback.

Reinforcement learning is fundamentally different: the learner is an *agent* that interacts with an *environment*. It learns by performing actions and receiving feedback in the form of *rewards*. The goal is to learn a strategy that maximizes the expected cumulative reward. Examples are playing games (e.g., AlphaGo), robotics control, and autonomous driving. Reinforcement learning corresponds to *learning by interaction*: the agent improves its behavior based on trial and error.

Algorithms in deep supervised learning often seem to be most accessible for a mathematical analysis. In the following we briefly sketch in a simplified situation some ideas of deep supervised learning. Let $d, M \in \mathbb{N} = \{1, 2, 3, \dots\}$, $\mathcal{E} \in C(\mathbb{R}^d, \mathbb{R})$, $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$ satisfy for all $m \in \{1, 2, \dots, M\}$ that

$$y_m = \mathcal{E}(x_m). \tag{1}$$

We think of $M \in \mathbb{N}$ as the number of available known input-output data pairs, we think of $d \in \mathbb{N}$ as the dimension of the input data, we think of $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ as an unknown function which relates input and output data through (1), we think of $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$ as the available known input data, and we think of $y_1, y_2, \dots, y_M \in \mathbb{R}$ as the available known output data.

In the context of a learning problem of the type (1) the objective then is to approximately compute the output $\mathcal{E}(x_{M+1})$ of the $(M + 1)$ -th input data x_{M+1} without using explicit

knowledge of the function $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ but instead by using the knowledge of the M input-output data pairs

$$(\mathbf{x}_1, \mathbf{y}_1) = (\mathbf{x}_1, \mathcal{E}(\mathbf{x}_1)), (\mathbf{x}_2, \mathbf{y}_2) = (\mathbf{x}_2, \mathcal{E}(\mathbf{x}_2)), \dots, (\mathbf{x}_M, \mathbf{y}_M) = (\mathbf{x}_M, \mathcal{E}(\mathbf{x}_M)) \in \mathbb{R}^d \times \mathbb{R}. \quad (2)$$

To accomplish this, one considers the optimization problem of computing approximate minimizers of the function $\mathfrak{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$ which satisfies for all $\phi \in C(\mathbb{R}^d, \mathbb{R})$ that

$$\mathfrak{L}(\phi) = \frac{1}{M} \left[\sum_{m=1}^M |\phi(\mathbf{x}_m) - \mathbf{y}_m|^2 \right]. \quad (3)$$

Observe that (1) ensures that $\mathfrak{L}(\mathcal{E}) = 0$ and, in particular, we have that the unknown function $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ in (1) above is a minimizer of the function

$$\mathfrak{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty). \quad (4)$$

The optimization problem of computing approximate minimizers of the function \mathfrak{L} is not suitable for discrete numerical computations on a computer as the function \mathfrak{L} is defined on the infinite dimensional vector space $C(\mathbb{R}^d, \mathbb{R})$.

To overcome this we introduce a spatially discretized version of this optimization problem. More specifically, let $n \in \mathbb{N}$, let $\psi = (\psi_\theta)_{\theta \in \mathbb{R}^n}: \mathbb{R}^n \rightarrow C(\mathbb{R}^d, \mathbb{R})$ be a function, and let $\mathcal{L}: \mathbb{R}^n \rightarrow [0, \infty)$ satisfy

$$\mathcal{L} = \mathfrak{L} \circ \psi. \quad (5)$$

We think of the set

$$\{\psi_\theta: \theta \in \mathbb{R}^n\} \subseteq C(\mathbb{R}^d, \mathbb{R}) \quad (6)$$

as a parametrized set of functions which we employ to approximate the infinite dimensional vector space $C(\mathbb{R}^d, \mathbb{R})$ and we think of the function

$$\mathbb{R}^n \ni \theta \mapsto \psi_\theta \in C(\mathbb{R}^d, \mathbb{R}) \quad (7)$$

as the parametrization function associated to this set. For example, in the case $d = 1$ one could think of (7) as the parametrization function associated to polynomials in the sense that for all $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$, $x \in \mathbb{R}$ it holds that

$$\psi_\theta(x) = \sum_{k=0}^{n-1} \theta_{k+1} x^k \quad (8)$$

or one could think of (7) as the parametrization associated to trigonometric polynomials. However, in the context of *deep supervised learning* one neither chooses (7) as parametrization

of polynomials nor as parametrization of trigonometric polynomials, but instead one chooses (7) as a parametrization associated to *deep* ANNs. In Chapter 1 in Part I we present different types of such deep ANN parametrization functions in all mathematical details.

Taking the set in (6) and its parametrization function in (7) into account, we then intend to compute approximate minimizers of the function \mathfrak{L} restricted to the set $\{\psi_\theta: \theta \in \mathbb{R}^n\}$, that is, we consider the optimization problem of computing approximate minimizers of the function

$$\{\psi_\theta: \theta \in \mathbb{R}^n\} \ni \phi \mapsto \mathfrak{L}(\phi) = \frac{1}{M} \left[\sum_{m=1}^M |\phi(\mathbf{x}_m) - \mathbf{y}_m|^2 \right] \in [0, \infty). \quad (9)$$

Employing the parametrization function in (7), one can also reformulate the optimization problem in (9) as the optimization problem of computing approximate minimizers of the function

$$\mathbb{R}^n \ni \theta \mapsto \mathcal{L}(\theta) = \mathfrak{L}(\psi_\theta) = \frac{1}{M} \left[\sum_{m=1}^M |\psi_\theta(\mathbf{x}_m) - \mathbf{y}_m|^2 \right] \in [0, \infty) \quad (10)$$

and this optimization problem now has the potential to be amenable for discrete numerical computations. In the context of deep supervised learning, where one chooses the parametrization function in (7) as deep ANN parametrizations, one would apply an SGD-type optimization algorithm to the optimization problem in (10) to compute approximate minimizers of (10). In Chapter 6 in Part III we present the most common variants of such SGD-type optimization algorithms. If $\vartheta \in \mathbb{R}^n$ is an approximate minimizer of (10) in the sense that $\mathcal{L}(\vartheta) \approx \inf_{\theta \in \mathbb{R}^n} \mathcal{L}(\theta)$, one then considers $\psi_{\vartheta}(\mathbf{x}_{M+1})$ as an approximation

$$\psi_{\vartheta}(\mathbf{x}_{M+1}) \approx \mathcal{E}(\mathbf{x}_{M+1}) \quad (11)$$

of the unknown output $\mathcal{E}(\mathbf{x}_{M+1})$ of the $(M+1)$ -th input data \mathbf{x}_{M+1} . We note that in deep supervised learning algorithms one typically aims to compute an approximate minimizer $\vartheta \in \mathbb{R}^n$ of (10) in the sense that $\mathcal{L}(\vartheta) \approx \inf_{\theta \in \mathbb{R}^n} \mathcal{L}(\theta)$, which is, however, typically not a minimizer of (10) in the sense that $\mathcal{L}(\vartheta) = \inf_{\theta \in \mathbb{R}^n} \mathcal{L}(\theta)$.

In (3) above we have set up an optimization problem for the learning problem by using the standard mean squared error function to measure the loss. This *mean squared error loss function* is just one possible example in the formulation of deep learning optimization problems. In particular, in image classification problems other loss functions such as the *cross-entropy loss function* are often used and we refer to Chapter 4 of Part III for a survey of commonly used loss function in deep learning algorithms (see Section 4.2.2).

The main questions concerning deep learning can roughly be categorized as follows:

- Expressivity: how good can ANNs approximate given family of functions?
- Learning: Why and when the training algorithm yields reasonable results?
- Generalization: Why do ANNs preform well on unseen data?

- Interpretability: Why did a trained ANN reach a certain decision?

Each of these categories is an extensive field of research with many scientific papers and practical simulations appearing in the fast few years. In this lecture, we will consider some of the aspects of *expressivity* and *learning*.

Part I

Artificial neural networks (ANNs)

Chapter 1

Basics on ANNs

In this chapter we review fully-connected feedforward ANNs (see Sections 1.1 and 1.3), different types of popular activation functions used in applications such as the *rectified linear unit* (ReLU) activation (see Section 1.2.3) among others, and different procedures for how ANNs can be formulated in rigorous mathematical terms (see Section 1.1 for a vectorized description and Section 1.3 for a structured description).

1.1 Fully-connected feedforward ANNs (vectorized description)

Roughly speaking, fully-connected feedforward ANNs can be thought of as parametric functions resulting from successive compositions of affine functions followed by nonlinear functions, where the parameters of a fully-connected feedforward ANN correspond to all the entries of the linear transformation matrices and translation vectors of the involved affine functions (cf.¹ Definition 1.1.3 below for a precise definition of fully-connected feedforward ANNs and Figure 1.2 below for a graphical illustration of fully-connected feedforward ANNs).

The linear transformation matrices and translation vectors are sometimes called *weight matrices* and *bias vectors*, respectively, and can be thought of as the *trainable parameters* of fully-connected feedforward ANNs (cf. Remark 1.1.6 below).

We note that there are more different type of ANNs employed and studied in the practice as well as in the literature, the most prominent ones being convolutional ANNs, residual ANNs, recurrent ANNs, ANNs with encoder-decoder architectures: autoencoders, transformers, graph neural networks, and neural operators. In these lecture notes we will

¹The abbreviation cf. (from Latin “confer” or “conferatur”, both meaning “compare”) is used in academic writing to refer the reader to other material to make a comparison with the topic being discussed. In these lecture notes we mostly use it to refer to definitions and mathematical results needed to comprehension of the considered notion/result.

only concentrate to fully-connected feedforward ANNs and, for the sake of conciseness, we just call them ANNs.

In this section we introduce in Definition 1.1.3 below a *vectorized description* of ANNs in the sense that all the trainable parameters of an ANN are represented by the components of a single Euclidean vector. In Section 1.3 below we will discuss an alternative way to describe ANNs in which the trainable parameters of an ANN are represented by a tuple of matrix-vector pairs corresponding to the weight matrices and bias vectors of ANNs (cf. Definitions 1.3.1 and 1.3.4 below).

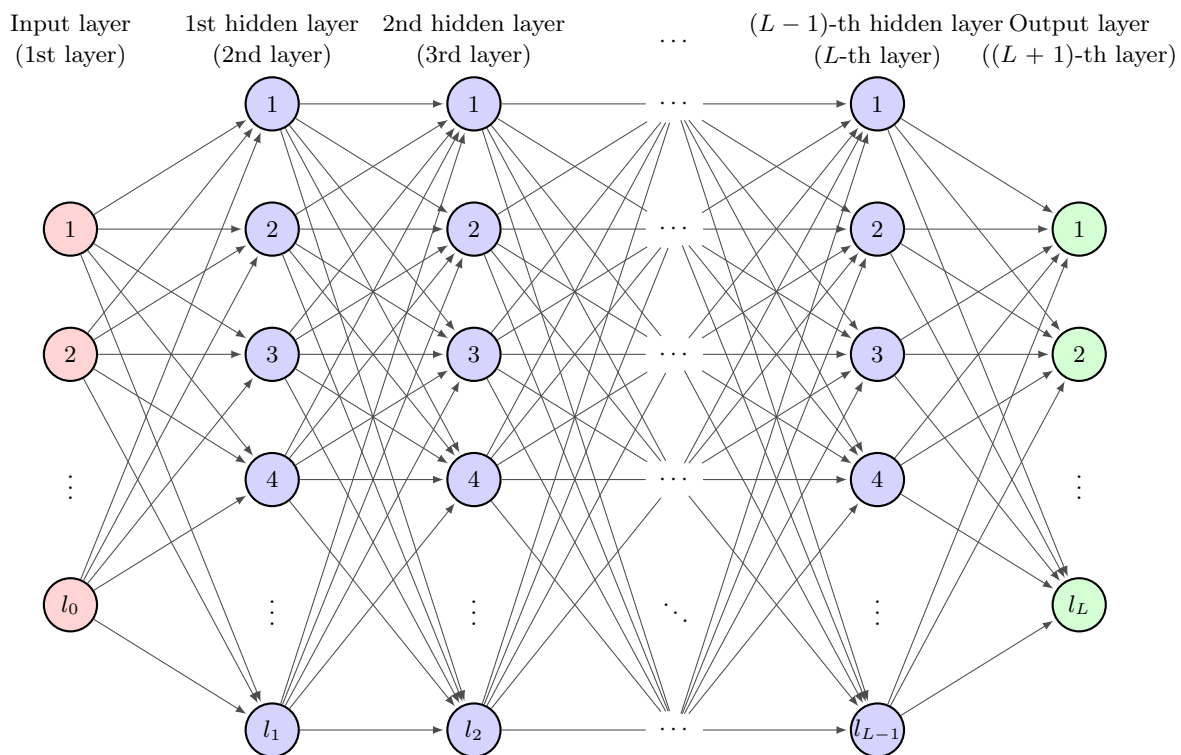


Figure 1.1: Graphical illustration of an ANN consisting of $L \in \mathbb{N}$ affine transformations (i.e., consisting of $L + 1$ layers: one input layer, $L - 1$ hidden layers, and one output layer) with $l_0 \in \mathbb{N}$ neurons on the input layer (i.e., with l_0 -dimensional input layer), with $l_1 \in \mathbb{N}$ neurons on the first hidden layer (i.e., with l_1 -dimensional first hidden layer), with $l_2 \in \mathbb{N}$ neurons on the second hidden layer (i.e., with l_2 -dimensional second hidden layer), ..., with l_{L-1} neurons on the $(L - 1)$ -th hidden layer (i.e., with (l_{L-1}) -dimensional $(L - 1)$ -th hidden layer), and with l_L neurons in the output layer (i.e., with l_L -dimensional output layer).

1.1.1 Affine functions

Definition 1.1.1 (Affine functions). Let $\mathfrak{d}, m, n \in \mathbb{N}$, $s \in \mathbb{N}_0$, $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathfrak{d} \geq s + mn + m$. Then we denote by $\mathcal{A}_{m,n}^{\theta,s} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ the function which satisfies for all $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ that

$$\begin{aligned} \mathcal{A}_{m,n}^{\theta,s}(x) &= \begin{pmatrix} \theta_{s+1} & \theta_{s+2} & \cdots & \theta_{s+n} \\ \theta_{s+n+1} & \theta_{s+n+2} & \cdots & \theta_{s+2n} \\ \theta_{s+2n+1} & \theta_{s+2n+2} & \cdots & \theta_{s+3n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{s+(m-1)n+1} & \theta_{s+(m-1)n+2} & \cdots & \theta_{s+mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \theta_{s+mn+1} \\ \theta_{s+mn+2} \\ \theta_{s+mn+3} \\ \vdots \\ \theta_{s+mn+m} \end{pmatrix} \\ &= \left(\left[\sum_{k=1}^n x_k \theta_{s+k} \right] + \theta_{s+mn+1}, \left[\sum_{k=1}^n x_k \theta_{s+n+k} \right] + \theta_{s+mn+2}, \dots, \right. \\ &\quad \left. \left[\sum_{k=1}^n x_k \theta_{s+(m-1)n+k} \right] + \theta_{s+mn+m} \right) \end{aligned} \quad (1.1)$$

and we call $\mathcal{A}_{m,n}^{\theta,s}$ the **affine function** from \mathbb{R}^n to \mathbb{R}^m associated to (θ, s) .

Note that in the above definition of $\mathcal{A}_{m,n}^{\theta,s}$

- the vector θ stands for parameter vector used in the construction of the weight matrix and the bias vector of the affine function $\mathcal{A}_{m,n}^{\theta,s}$,
- $s + 1$ is the index of the parameter vector θ from which the parameters are actually used in the construction $\mathcal{A}_{m,n}^{\theta,s}$,
- m is the output dimension and n is the input dimension of $\mathcal{A}_{m,n}^{\theta,s}$.

Question: How many parameters are used in the construction of $\mathcal{A}_{m,n}^{\theta,s}$?

Example 1.1.2 (Example for Definition 1.1.1). Let $\theta = (0, 1, 2, 0, 3, 3, 0, 1, 7) \in \mathbb{R}^9$. Then

$$\mathcal{A}_{2,2}^{\theta,1}((1, 2)) = (8, 6) \quad (1.2)$$

(cf. Definition 1.1.1).

Proof for Example 1.1.2. Observe that (1.1) ensures that

$$\mathcal{A}_{2,2}^{\theta,1}((1, 2)) = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 1+4 \\ 0+6 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \end{pmatrix}. \quad (1.3)$$

The proof for Example 1.1.2 is thus complete. \square

Exercise 1.1.1. Let $\theta = (3, 1, -2, 1, -3, 0, 5, 4, -1, -1, 0) \in \mathbb{R}^{11}$. Specify $\mathcal{A}_{2,3}^{\theta,2}((-1, 1, -1))$ explicitly and prove that your result is correct (cf. Definition 1.1.1).

1.1.2 Vectorized description of ANNs

One can now use the above defined affine functions to construct realizations functions associated to ANNs. We will first restrict ourselves to vectorized description of ANNs, meaning that the parameters involved in the construction of the all used affine functions are stored in a single vector θ . Later in the lecture notes we will also consider the structured description of ANNs and the relation of the both descriptions.

Definition 1.1.3 (Vectorized description of ANNs). Let $\mathfrak{d}, L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\mathfrak{d} \geq \sum_{k=1}^L l_k(l_{k-1} + 1) \quad (1.4)$$

and for every $k \in \{1, 2, \dots, L\}$ let $\Psi_k: \mathbb{R}^{l_k} \rightarrow \mathbb{R}^{l_k}$ be a function. Then we denote by $\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0}: \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L}$ the function which satisfies for all $x \in \mathbb{R}^{l_0}$ that

$$\begin{aligned} (\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x) = & (\Psi_L \circ \mathcal{A}_{l_L, l_{L-1}}^{\theta, \sum_{k=1}^{L-1} l_k(l_{k-1}+1)} \circ \Psi_{L-1} \circ \mathcal{A}_{l_{L-1}, l_{L-2}}^{\theta, \sum_{k=1}^{L-2} l_k(l_{k-1}+1)} \circ \dots \\ & \dots \circ \Psi_2 \circ \mathcal{A}_{l_2, l_1}^{\theta, l_1(l_0+1)} \circ \Psi_1 \circ \mathcal{A}_{l_1, l_0}^{\theta, 0})(x) \end{aligned} \quad (1.5)$$

and we call $\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0}$ the **realization function** of the ANN associated to θ with $L + 1$ layers with dimensions (l_0, l_1, \dots, l_L) and activation functions $(\Psi_1, \Psi_2, \dots, \Psi_L)$ (cf. Definition 1.1.1).

Note that in the above definition the realization function $\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0}$ is constructed via consecutive compositions of affine (i.e. linear) functions with possibly nonlinear functions $(\Psi_1, \Psi_2, \dots, \Psi_L)$. More precisely, first, to go from layer 1 to layer 2 we apply

$$\Psi_1 \circ \mathcal{A}_{l_1, l_0}^{\theta, 0}. \quad (1.6)$$

Since $\mathcal{A}_{l_1, l_0}^{\theta, 0}$ uses the first $l_1(l_0 + 1)$ entries of θ , on layer 2 we employ

$$\Psi_2 \circ \mathcal{A}_{l_2, l_1}^{\theta, l_1(l_0+1)} \quad (1.7)$$

to go to layer 3, and so on.

Remark 1.1.4. In practice one often sets the last activation function Ψ_L to be equal to the identity, i.e. there is no activation applied on the last layer.

Example 1.1.5 (Example for Definition 1.1.3). Let $\theta = (1, -1, 2, -2, 3, -3, 0, 0, 1) \in \mathbb{R}^9$ and let $\Psi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ satisfy for all $x = (x_1, x_2) \in \mathbb{R}^2$ that

$$\Psi(x) = (\max\{x_1, 0\}, \max\{x_2, 0\}). \quad (1.8)$$

Then

$$(\mathcal{N}_{\Psi, \text{id}_{\mathbb{R}}}^{\theta, 1})(2) = 12 \quad (1.9)$$

(cf. Definition 1.1.3).

Proof for Example 1.1.5. Note that (1.1), (1.5), and (1.8) assure that

$$\begin{aligned} (\mathcal{N}_{\Psi, \text{id}_{\mathbb{R}}}^{\theta, 1})(2) &= (\text{id}_{\mathbb{R}} \circ \mathcal{A}_{1,2}^{\theta, 4} \circ \Psi \circ \mathcal{A}_{2,1}^{\theta, 0})(2) = (\mathcal{A}_{1,2}^{\theta, 4} \circ \Psi) \left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}(2) + \begin{pmatrix} 2 \\ -2 \end{pmatrix} \right) \\ &= (\mathcal{A}_{1,2}^{\theta, 4} \circ \Psi) \left(\begin{pmatrix} 4 \\ -4 \end{pmatrix} \right) = \mathcal{A}_{1,2}^{\theta, 4} \left(\begin{pmatrix} 4 \\ 0 \end{pmatrix} \right) = (3 \quad -3) \begin{pmatrix} 4 \\ 0 \end{pmatrix} + (0) = 12. \end{aligned} \quad (1.10)$$

The proof for Example 1.1.5 is thus complete. \square

Exercise 1.1.2. Let $\theta = (1, -1, 0, 0, 1, -1, 0) \in \mathbb{R}^7$ and let $\Psi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ satisfy for all $x = (x_1, x_2) \in \mathbb{R}^2$ that

$$\Psi(x) = (\max\{x_1, 0\}, \min\{x_2, 0\}). \quad (1.11)$$

Prove or disprove the following statement: It holds that

$$(\mathcal{N}_{\Psi, \text{id}_{\mathbb{R}}}^{\theta, 1})(-1) = -1 \quad (1.12)$$

(cf. Definition 1.1.3).

Exercise 1.1.3. Let $\theta = (\theta_1, \theta_2, \dots, \theta_{10}) \in \mathbb{R}^{10}$ satisfy

$$\theta = (\theta_1, \theta_2, \dots, \theta_{10}) = (1, 0, 2, -1, 2, 0, -1, 1, 2, 1)$$

and let $m: \mathbb{R} \rightarrow \mathbb{R}$ and $q: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that

$$m(x) = \max\{-x, 0\} \quad \text{and} \quad q(x) = x^2. \quad (1.13)$$

Specify $(\mathcal{N}_{q, m, q}^{\theta, 1})(0)$, $(\mathcal{N}_{q, m, q}^{\theta, 1})(1)$, and $(\mathcal{N}_{q, m, q}^{\theta, 1})(1/2)$ explicitly and prove that your results are correct (cf. Definition 1.1.3).

Exercise 1.1.4. Let $\theta = (\theta_1, \theta_2, \dots, \theta_{15}) \in \mathbb{R}^{15}$ satisfy

$$(\theta_1, \theta_2, \dots, \theta_{15}) = (1, -2, 0, 3, 2, -1, 0, 3, 1, -1, 1, -1, 2, 0, -1) \quad (1.14)$$

and let $\Phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ and $\Psi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ satisfy for all $x, y \in \mathbb{R}$ that $\Phi(x, y) = (y, x)$ and $\Psi(x, y) = (xy, xy)$.

- a) Prove or disprove the following statement: It holds that $(\mathcal{N}_{\Phi, \Psi}^{\theta, 2})(1, -1) = (4, 4)$ (cf. Definition 1.1.3).
- b) Prove or disprove the following statement: It holds that $(\mathcal{N}_{\Phi, \Psi}^{\theta, 2})(-1, 1) = (-4, -4)$ (cf. Definition 1.1.3).

1.1.3 Weight and bias parameters of ANNs

Remark 1.1.6 (Weights and biases for ANNs). Let $L \in \{2, 3, 4, \dots\}$, $v_0, v_1, \dots, v_{L-1} \in \mathbb{N}_0$, $l_0, l_1, \dots, l_L, \mathfrak{d} \in \mathbb{N}$, $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $k \in \{0, 1, \dots, L-1\}$ that

$$\mathfrak{d} \geq \sum_{i=1}^L l_i(l_{i-1} + 1) \quad \text{and} \quad v_k = \sum_{i=1}^k l_i(l_{i-1} + 1), \quad (1.15)$$

let $W_k \in \mathbb{R}^{l_k \times l_{k-1}}$, $k \in \{1, 2, \dots, L\}$, and $b_k \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \dots, L\}$, satisfy for all $k \in \{1, 2, \dots, L\}$ that

$$W_k = \begin{pmatrix} \theta_{v_{k-1}+1} & \theta_{v_{k-1}+2} & \cdots & \theta_{v_{k-1}+l_{k-1}} \\ \theta_{v_{k-1}+l_{k-1}+1} & \theta_{v_{k-1}+l_{k-1}+2} & \cdots & \theta_{v_{k-1}+2l_{k-1}} \\ \theta_{v_{k-1}+2l_{k-1}+1} & \theta_{v_{k-1}+2l_{k-1}+2} & \cdots & \theta_{v_{k-1}+3l_{k-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{v_{k-1}+(l_{k-1})l_{k-1}+1} & \theta_{v_{k-1}+(l_{k-1})l_{k-1}+2} & \cdots & \theta_{v_{k-1}+l_k l_{k-1}} \end{pmatrix} \quad (1.16)$$

weight parameters

$$\text{and} \quad b_k = \underbrace{(\theta_{v_{k-1}+l_k l_{k-1}+1}, \theta_{v_{k-1}+l_k l_{k-1}+2}, \dots, \theta_{v_{k-1}+l_k l_{k-1}+l_k})}_{\text{bias parameters}}, \quad (1.17)$$

and let $\Psi_k: \mathbb{R}^{l_k} \rightarrow \mathbb{R}^{l_k}$, $k \in \{1, 2, \dots, L\}$, be functions. Then

- (i) it holds that

$$\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0} = \Psi_L \circ \mathcal{A}_{l_L, l_{L-1}}^{\theta, v_{L-1}} \circ \Psi_{L-1} \circ \mathcal{A}_{l_{L-1}, l_{L-2}}^{\theta, v_{L-2}} \circ \Psi_{L-2} \circ \cdots \circ \mathcal{A}_{l_2, l_1}^{\theta, v_1} \circ \Psi_1 \circ \mathcal{A}_{l_1, l_0}^{\theta, v_0} \quad (1.18)$$

and

- (ii) it holds for all $k \in \{1, 2, \dots, L\}$, $x \in \mathbb{R}^{l_{k-1}}$ that $\mathcal{A}_{l_k, l_{k-1}}^{\theta, v_{k-1}}(x) = W_k x + b_k$

(cf. Definitions 1.1.1 and 1.1.3). Hence, for every $k \in \{1, 2, \dots, L\}$ we have that W_k is the *weight matrix* and b_k is the *bias vector* associated to the k -th affine linear transformation.

Question: In the above remark what is v_0 ? How can one describe v_k ?

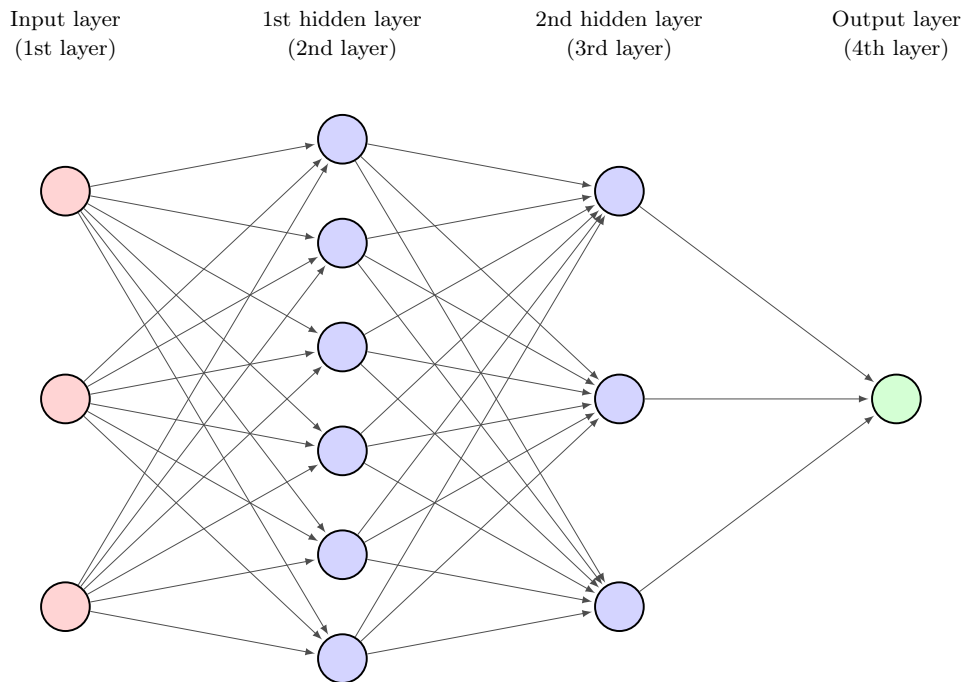


Figure 1.2: Graphical illustration of an ANN. The ANN has 2 hidden layers and length $L = 3$ with 3 neurons in the input layer (corresponding to $l_0 = 3$), 6 neurons in the first hidden layer (corresponding to $l_1 = 6$), 3 neurons in the second hidden layer (corresponding to $l_2 = 3$), and one neuron in the output layer (corresponding to $l_3 = 1$). In this situation we have an ANN with 39 weight parameters and 10 bias parameters adding up to 49 parameters overall. The realization of this ANN is a function from \mathbb{R}^3 to \mathbb{R} .

1.2 Activation functions

In this section we review a few popular activation functions from the literature.

1.2.1 Multidimensional versions

To describe multidimensional activation functions, we frequently employ the concept of the multidimensional version of a function. This concept is the subject of the next notion.

Definition 1.2.1 (Multidimensional versions of one-dimensional functions). *Let $d \in \mathbb{N}$ and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then we denote by*

$$\mathfrak{M}_{\psi,d}: \mathbb{R}^d \rightarrow \mathbb{R}^d \quad (1.19)$$

the function which satisfies for all $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ that

$$\mathfrak{M}_{\psi,d}(x) = (\psi(x_1), \psi(x_2), \dots, \psi(x_d)) \quad (1.20)$$

*and we call $\mathfrak{M}_{\psi,d}$ the **d -dimensional version** of ψ .*

Example 1.2.2 (Example for Definition 1.2.1). *Let $y = (1, -2, -3) \in \mathbb{R}^3$ and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that $\psi(x) = x^2$. Then*

$$\mathfrak{M}_{\psi,3}(y) = (1, 4, 9). \quad (1.21)$$

Exercise 1.2.1. Let $y \in \mathbb{R}^3$, $z \in \mathbb{R}^2$ satisfy

$$y = (3, -2, 5) \quad \text{and} \quad z = (-3, 5) \quad (1.22)$$

and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that $\psi(x) = |x|$. Specify $\mathfrak{M}_{\psi,3}(y)$ and $\mathfrak{M}_{\psi,2}(z)$ explicitly and prove that your results are correct (cf. Definition 1.2.1).

Exercise 1.2.2. Let $\theta = (\theta_1, \theta_2, \dots, \theta_{14}) \in \mathbb{R}^{14}$ satisfy

$$(\theta_1, \theta_2, \dots, \theta_{14}) = (0, 1, 2, 2, 1, 0, 1, 1, 1, -3, -1, 4, 0, 1) \quad (1.23)$$

and let $f: \mathbb{R} \rightarrow \mathbb{R}$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that

$$f(x) = \frac{1}{1 + |x|} \quad \text{and} \quad g(x) = x^2. \quad (1.24)$$

Specify $(\mathcal{N}_{\mathfrak{M}_{f,3}, \mathfrak{M}_{g,2}}^{\theta,1})(1)$ and $(\mathcal{N}_{\mathfrak{M}_{g,2}, \mathfrak{M}_{f,3}}^{\theta,1})(1)$ explicitly and prove that your results are correct (cf. Definitions 1.1.3 and 1.2.1).

1.2.2 Single hidden layer ANNs

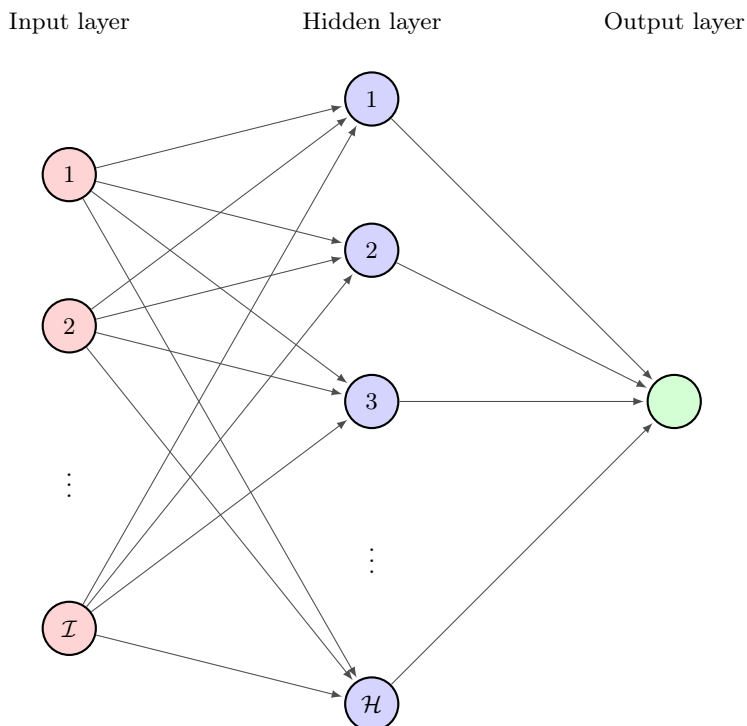


Figure 1.3: Graphical illustration of an ANN consisting of two affine transformations (i.e., consisting of 3 layers: one input layer, one hidden layer, and one output layer) with $\mathcal{I} \in \mathbb{N}$ neurons on the input layer (i.e., with \mathcal{I} -dimensional input layer), with $\mathcal{H} \in \mathbb{N}$ neurons on the hidden layer (i.e., with \mathcal{H} -dimensional hidden layer), and with one neuron in the output layer (i.e., with 1-dimensional output layer).

Lemma 1.2.3 (ANN with one hidden layer). *Let $\mathcal{I}, \mathcal{H} \in \mathbb{N}$, $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}) \in \mathbb{R}^{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}$, $x = (x_1, x_2, \dots, x_{\mathcal{I}}) \in \mathbb{R}^{\mathcal{I}}$ and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then*

$$\mathcal{N}_{\mathfrak{M}_{\psi, \mathcal{H}, \text{id}_{\mathbb{R}}}}^{\theta, \mathcal{I}}(x) = \left[\sum_{k=1}^{\mathcal{H}} \theta_{\mathcal{H}\mathcal{I}+\mathcal{H}+k} \psi \left(\left[\sum_{i=1}^{\mathcal{I}} x_i \theta_{(k-1)\mathcal{I}+i} \right] + \theta_{\mathcal{H}\mathcal{I}+k} \right) \right] + \theta_{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}. \quad (1.25)$$

(cf. Definitions 1.1.1, 1.1.3, and 1.2.1).

Proof of Lemma 1.2.3. Observe that (1.5) and (1.20) show that

$$\begin{aligned}
 & \mathcal{N}_{\mathfrak{M}_{\psi, \mathcal{H}}, \text{id}_{\mathbb{R}}}^{\theta, \mathcal{I}}(x) \\
 &= \left(\text{id}_{\mathbb{R}} \circ \mathcal{A}_{1, \mathcal{H}}^{\theta, \mathcal{H}\mathcal{I} + \mathcal{H}} \circ \mathfrak{M}_{\psi, \mathcal{H}} \circ \mathcal{A}_{\mathcal{H}, \mathcal{I}}^{\theta, 0} \right)(x) \\
 &= \mathcal{A}_{1, \mathcal{H}}^{\theta, \mathcal{H}\mathcal{I} + \mathcal{H}} \left(\mathfrak{M}_{\psi, \mathcal{H}} \left(\mathcal{A}_{\mathcal{H}, \mathcal{I}}^{\theta, 0}(x) \right) \right) \\
 &= \left[\sum_{k=1}^{\mathcal{H}} \theta_{\mathcal{H}\mathcal{I} + \mathcal{H} + k} \psi \left(\left[\sum_{i=1}^{\mathcal{I}} x_i \theta_{(k-1)\mathcal{I} + i} \right] + \theta_{\mathcal{H}\mathcal{I} + k} \right) \right] + \theta_{\mathcal{H}\mathcal{I} + 2\mathcal{H} + 1}.
 \end{aligned} \tag{1.26}$$

The proof of Lemma 1.2.3 is thus complete. \square

1.2.3 Rectified linear unit (ReLU) activation

In this subsection we formulate the ReLU function which is one of the most frequently used activation functions in deep learning applications (cf., for example, LeCun et al. [5]).

Definition 1.2.4 (ReLU activation function). *We denote by $\mathfrak{r}: \mathbb{R} \rightarrow \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that*

$$\mathfrak{r}(x) = \max\{x, 0\} \tag{1.27}$$

*and we call \mathfrak{r} the **ReLU** activation function (we call \mathfrak{r} the **rectifier** function).*

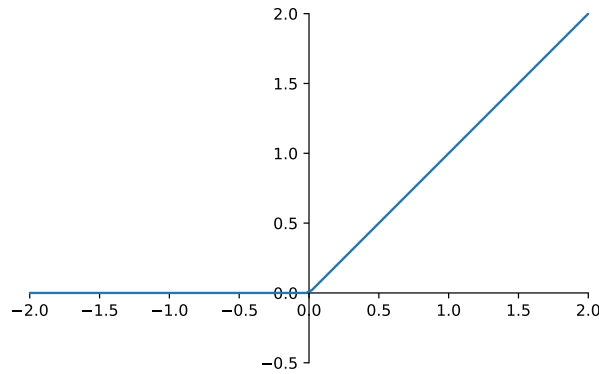


Figure 1.4 ([plots/relu.pdf](#)): A plot of the ReLU activation function.

Definition 1.2.5 (Multidimensional ReLU activation functions). *Let $d \in \mathbb{N}$. Then we denote by $\mathfrak{R}_d: \mathbb{R}^d \rightarrow \mathbb{R}^d$ the function given by*

$$\mathfrak{R}_d = \mathfrak{M}_{\mathfrak{r}, d} \tag{1.28}$$

and we call \mathfrak{R}_d the d -dimensional ReLU activation function (we call \mathfrak{R}_d the d -dimensional rectifier function) (cf. Definitions 1.2.1 and 1.2.4).

Lemma 1.2.6 (An ANN with the ReLU activation function as the activation function). Let $W_1 = w_1 = 1$, $W_2 = w_2 = -1$, $b_1 = b_2 = B = 0$. Then it holds for all $x \in \mathbb{R}$ that

$$x = W_1 \max\{w_1 x + b_1, 0\} + W_2 \max\{w_2 x + b_2, 0\} + B. \quad (1.29)$$

Proof of Lemma 1.2.6. Observe that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} & W_1 \max\{w_1 x + b_1, 0\} + W_2 \max\{w_2 x + b_2, 0\} + B \\ &= \max\{w_1 x + b_1, 0\} - \max\{w_2 x + b_2, 0\} = \max\{x, 0\} - \max\{-x, 0\} \\ &= \max\{x, 0\} + \min\{x, 0\} = x. \end{aligned} \quad (1.30)$$

The proof of Lemma 1.2.6 is thus complete. \square

Lemma 1.2.7 follows from an application of Lemma 1.2.6 and the detailed proof of Lemma 1.2.7 is left as an exercise. It presents an important property of the ReLU activation function, namely, that an ANN with the ReLU activation function is able to realize the identity function.

Lemma 1.2.7 (Real identity). Let $\theta = (1, -1, 0, 0, 1, -1, 0) \in \mathbb{R}^7$. Then it holds for all $x \in \mathbb{R}$ that

$$(\mathcal{N}_{\mathfrak{R}_2, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.31)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.3 (Absolute value). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + l_H + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = |x| \quad (1.32)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.4 (Exponential). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + l_H + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = e^x \quad (1.33)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.5 (Two-dimensional maximum). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 3l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + l_H + 1$ such that for all $x, y \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 2})(x, y) = \max\{x, y\} \quad (1.34)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.6 (Real identity with two hidden layers). Prove or disprove the following statement: There exist $\mathfrak{d}, l_1, l_2 \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + l_1l_2 + 2l_2 + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.35)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.7 (Three-dimensional maximum). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 4l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + l_H + 1$ such that for all $x, y, z \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 3})(x, y, z) = \max\{x, y, z\} \quad (1.36)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.8 (Multidimensional maxima). Prove or disprove the following statement: For every $k \in \mathbb{N}$ there exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq (k + 1)l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + l_H + 1$ such that for all $x_1, x_2, \dots, x_k \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, k})(x_1, x_2, \dots, x_k) = \max\{x_1, x_2, \dots, x_k\} \quad (1.37)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.9. Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + (l_H + 1)$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \max\{x, \frac{x}{2}\} \quad (1.38)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.10 (Hat function). Prove or disprove the following statement: There exist $\mathfrak{d}, l \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 3l + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_l, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \begin{cases} 1 & : x \leq 2 \\ x - 1 & : 2 < x \leq 3 \\ 5 - x & : 3 < x \leq 4 \\ 1 & : x > 4 \end{cases} \quad (1.39)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.11. Prove or disprove the following statement: There exist $\mathfrak{d}, l \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 3l + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_l, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \begin{cases} -2 & : x \leq 1 \\ 2x - 4 & : 1 < x \leq 3 \\ 2 & : x > 3 \end{cases} \quad (1.40)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.12. Prove or disprove the following statement: There exists $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + (l_H + 1)$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_{l_1, \mathfrak{R}_{l_2, \dots, \mathfrak{R}_{l_H}}, \text{id}_{\mathbb{R}}}}^{\theta, 1})(x) = \begin{cases} 0 & : x \leq 1 \\ x - 1 & : 1 \leq x \leq 2 \\ 1 & : x \geq 2 \end{cases} \quad (1.41)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.13. Prove or disprove the following statement: There exist $\mathfrak{d}, l \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 3l + 1$ such that for all $x \in [0, 1]$ it holds that

$$(\mathcal{N}_{\mathfrak{R}_l, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x^2 \quad (1.42)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.14. Prove or disprove the following statement: There exists $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + (l_H + 1)$ such that

$$\sup_{x \in [-3, -2]} |(\mathcal{N}_{\mathfrak{R}_{l_1, \mathfrak{R}_{l_2, \dots, \mathfrak{R}_{l_H}}, \text{id}_{\mathbb{R}}}}^{\theta, 1})(x) - (x + 2)^2| \leq \frac{1}{4} \quad (1.43)$$

(cf. Definitions 1.1.3 and 1.2.5).

1.2.4 Other activation functions

The below defined softplus activation function serves as a smooth approximation of the ReLU function. It is continuously differentiable, convex, and strictly increasing. Unlike ReLU, it avoids nondifferentiability at zero, ensuring smooth gradient flow while maintaining similar behavior—linear for large ($x > 0$) and nearly zero for large negative x .

Definition 1.2.8 (Softplus activation function). *We say that a is the softplus activation function if $a: \mathbb{R} \rightarrow \mathbb{R}$ satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \ln(1 + \exp(x)). \quad (1.44)$$

The next result, Lemma 1.2.9 below, presents a few elementary properties of the softplus function.

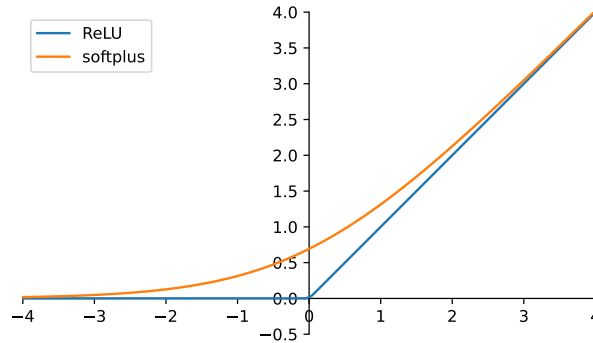


Figure 1.5 ([plots/softplus.pdf](#)): A plot of the softplus activation function and the ReLU activation function

Lemma 1.2.9 (Properties of the softplus function). *Let a be the softplus activation function (cf. Definition 1.2.8). Then*

- (i) *it holds for all $x \in [0, \infty)$ that $x \leq a(x) \leq x + 1$,*
- (ii) *it holds that $\lim_{x \rightarrow -\infty} a(x) = 0$,*
- (iii) *it holds that $\lim_{x \rightarrow \infty} a(x) = \infty$, and*
- (iv) *it holds that $a(0) = \ln(2)$.*

Proof of Lemma 1.2.9. Observe that for all $x \in [0, \infty)$ it holds that

$$\begin{aligned} x &= \ln(\exp(x)) \leq \ln(1 + \exp(x)) = \ln(\exp(0) + \exp(x)) \\ &\leq \ln(\exp(x) + \exp(x)) = \ln(2 \exp(x)). \end{aligned} \tag{1.45}$$

Combining with the fact that $2 \leq \exp(1)$ ensures that for all $x \in [0, \infty)$ it holds that

$$x \leq a(x) \leq \ln(\exp(1) \exp(x)) = \ln(\exp(x + 1)) = x + 1. \tag{1.46}$$

The proof of Lemma 1.2.9 is thus complete. \square

Definition 1.2.10 (Hyperbolic tangent activation function). *We denote by $\tanh: \mathbb{R} \rightarrow \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that*

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{1.47}$$

and we call \tanh the hyperbolic tangent activation function.

Definition 1.2.11 (Softsign activation function). We say that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the softsign activation function if it satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \frac{x}{|x| + 1}. \quad (1.48)$$

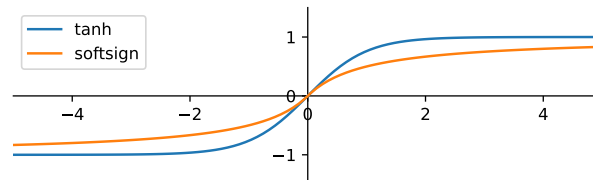


Figure 1.6 ([plots/softsign.pdf](#)): A plot of the softsign activation function and the hyperbolic tangent

Lemma 1.2.12 (Properties of the softsign activation). Let a be softsign activation function (cf. Definition 1.2.11). Then it holds that

(i) *Oddness and bounds: a is odd, i.e. $a(-x) = -a(x)$, and is bounded with*

$$-1 < a(x) < 1, \quad \lim_{x \rightarrow \pm\infty} a(x) = \pm 1.$$

(ii) *Monotonicity and differentiability: a is strictly increasing and continuously differentiable on \mathbb{R} with*

$$a'(x) = \frac{1}{(1 + |x|)^2} \in (0, 1].$$

(iii) *Lipschitz continuity:*

$$|a(x) - a(y)| \leq |x - y| \quad \text{for all } x, y \in \mathbb{R}.$$

Thus a is 1-Lipschitz.

(iv) *Smoothness of the derivative: a' is globally 2-Lipschitz, i.e.*

$$|a'(x) - a'(y)| \leq 2|x - y| \quad \text{for all } x, y \in \mathbb{R}.$$

Moreover, $a \in C^1(\mathbb{R})$, piecewise C^∞ , but $a \notin C^2(\mathbb{R})$ since the second derivative has a jump at 0.

(v) *Range and inverse: a is a bijection from \mathbb{R} onto $(-1, 1)$ with the inverse*

$$a^{-1}(y) = \frac{y}{1 - |y|}, \quad y \in (-1, 1).$$

The next activation function, leaky ReLU activation function, is piecewise linear like ReLU. It is continuous, monotone increasing, and piecewise differentiable with derivative equal to 1 for $x > 0$ and equal to γ for $x < 0$, ensuring nonvanishing gradients even for negative inputs – unlike the standard ReLU.

Definition 1.2.13 (Leaky ReLU activation function). *Let $\gamma \in [0, \infty)$. Then we say that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the leaky ReLU activation function with leak factor γ if it satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \begin{cases} x & : x > 0 \\ \gamma x & : x \leq 0. \end{cases} \quad (1.49)$$

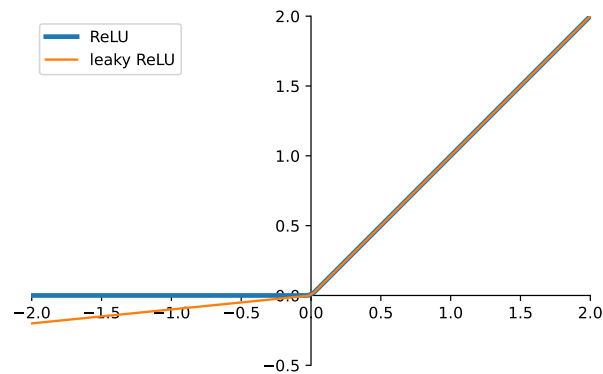


Figure 1.7 ([plots/leaky_relu.pdf](#)): A plot of the leaky ReLU activation function with leak factor $1/10$ and the ReLU activation function

Lemma 1.2.14. *Let $\gamma \in [0, 1]$ and let $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then a is the leaky ReLU activation function with leak factor γ for all $x \in \mathbb{R}$ we have*

$$a(x) = \max\{x, \gamma x\} \quad (1.50)$$

(cf. Definition 1.2.13).

Another popular activation function is the *Gaussian error linear unit* (GELU) activation function first introduced in Hendrycks & Gimpel [2]. It is defined via the cumulative distribution function of the standard normal distribution. It represents a smooth probabilistic gating mechanism, weighting each input by the probability that a standard normal variable is less than x , making it infinitely differentiable, non-linear, and approximately similar to ReLU for large $x > 0$ while retaining smoothness near the origin.

Definition 1.2.15 (GELU activation function). We say that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the GELU unit activation function (we say that a is the GELU activation function) if it satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \frac{x}{\sqrt{2\pi}} \left[\int_{-\infty}^x \exp\left(-\frac{z^2}{2}\right) dz \right]. \quad (1.51)$$

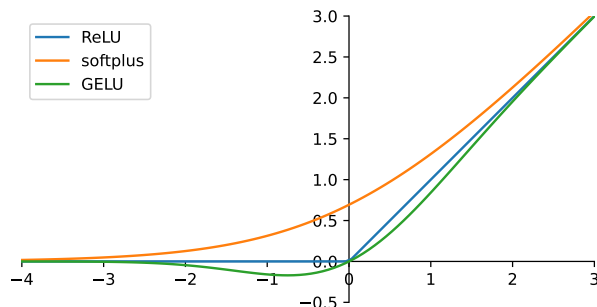


Figure 1.8 ([plots/gelu.pdf](#)): A plot of the GELU activation function, the ReLU activation function, and the softplus activation function

Lemma 1.2.16. Let $x \in \mathbb{R}$ and let a be the GELU activation function (cf. Definition 1.2.15). Then the following two statements are equivalent:

- (i) It holds that $a(x) > 0$.
- (ii) It holds that $\mathfrak{r}(x) > 0$ (cf. Definition 1.2.4).

The next activation function behaves linearly for large positive inputs and saturates near zero for large negative inputs. In addition it combines the favorable analytical properties of smoothness, bounded derivative, and approximate linearity with the non-monotonicity that enhances representational power.

Definition 1.2.17 (Swish activation function). Let $\beta \in \mathbb{R}$. Then we say that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the swish activation function with parameter β if it satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \frac{x}{1 + \exp(-\beta x)}. \quad (1.52)$$

Lemma 1.2.18. For every $\beta \in \mathbb{R}$ let $a_\beta: \mathbb{R} \rightarrow \mathbb{R}$ be the swish activation function with parameter β . Then

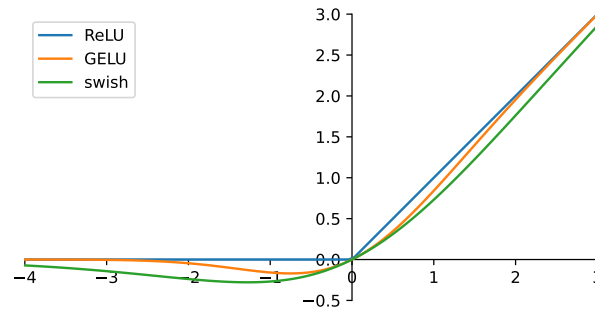


Figure 1.9 ([plots/swish.pdf](#)): A plot of the swish activation function, the GELU activation function, and the ReLU activation function

(i) for every $\beta \in \mathbb{R}$ the function a_β is infinitely often differentiable and

(ii) for every $x \in \mathbb{R}$

$$\lim_{\beta \rightarrow \infty} a_\beta(x) = \mathbf{r}(x). \quad (1.53)$$

(cf. Definition 1.2.4)

1.3 ANNs (structured description)

In this section we present an alternative way to describe the ANNs introduced in Section 1.1 above. Roughly speaking, in Section 1.1 above we defined a *vectorized description* of ANNs in the sense that the trainable parameters of an ANN are represented by the components of a single Euclidean vector (cf. Definition 1.1.3 above). In this section we introduce a *structured description* of ANNs in which the trainable parameters of an ANN are represented by a tuple of matrix-vector pairs corresponding to the weight matrices and bias vectors of the ANNs (cf. Definitions 1.3.1 and 1.3.4 below).

1.3.1 Structured description of ANNs

Definition 1.3.1 (Structured description of ANNs). We denote by \mathbf{N} the set given by

$$\mathbf{N} = \bigcup_{L \in \mathbb{N}} \bigcup_{l_0, l_1, \dots, l_L \in \mathbb{N}} \left(\prod_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right), \quad (1.54)$$

for every $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\Phi \in (\times_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \subseteq \mathbf{N}$ we denote by $\mathcal{P}(\Phi), \mathcal{L}(\Phi), \mathcal{I}(\Phi), \mathcal{O}(\Phi) \in \mathbb{N}$, $\mathcal{H}(\Phi) \in \mathbb{N}_0$ the numbers given by

$$\mathcal{P}(\Phi) = \sum_{k=1}^L l_k(l_{k-1} + 1), \quad \mathcal{L}(\Phi) = L, \quad \mathcal{I}(\Phi) = l_0, \quad \mathcal{O}(\Phi) = l_L, \quad \text{and} \quad \mathcal{H}(\Phi) = L - 1, \quad (1.55)$$

for every $n \in \mathbb{N}_0$, $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\Phi \in (\times_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \subseteq \mathbf{N}$ we denote by $\mathbb{D}_n(\Phi) \in \mathbb{N}_0$ the number given by

$$\mathbb{D}_n(\Phi) = \begin{cases} l_n & : n \leq L \\ 0 & : n > L, \end{cases} \quad (1.56)$$

for every $\Phi \in \mathbf{N}$ we denote by $\mathcal{D}(\Phi) \in \mathbb{N}^{\mathcal{L}(\Phi)+1}$ the tuple given by

$$\mathcal{D}(\Phi) = (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \quad (1.57)$$

and for every $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\Phi = ((W_1, B_1), \dots, (W_L, B_L)) \in (\times_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \subseteq \mathbf{N}$, $n \in \{1, 2, \dots, L\}$ we denote by $\mathcal{W}_{n,\Phi} \in \mathbb{R}^{l_n \times l_{n-1}}$, $\mathcal{B}_{n,\Phi} \in \mathbb{R}^{l_n}$ the matrix and the vector given by

$$\mathcal{W}_{n,\Phi} = W_n \quad \text{and} \quad \mathcal{B}_{n,\Phi} = B_n. \quad (1.58)$$

Definition 1.3.2 (ANNs). We say that Φ is an ANN if

$$\Phi \in \mathbf{N} \quad (1.59)$$

(cf. Definition 1.3.1).

Lemma 1.3.3. Let $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then

(i) it holds that $\mathcal{D}(\Phi) \in \mathbb{N}^{\mathcal{L}(\Phi)+1}$,

(ii) it holds that

$$\mathcal{I}(\Phi) = \mathbb{D}_0(\Phi) \quad \text{and} \quad \mathcal{O}(\Phi) = \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi), \quad (1.60)$$

and

(iii) it holds for all $n \in \{1, 2, \dots, \mathcal{L}(\Phi)\}$ that

$$\mathcal{W}_{n,\Phi} \in \mathbb{R}^{\mathbb{D}_n(\Phi) \times \mathbb{D}_{n-1}(\Phi)} \quad \text{and} \quad \mathcal{B}_{n,\Phi} \in \mathbb{R}^{\mathbb{D}_n(\Phi)}. \quad (1.61)$$

Proof of Lemma 1.3.3. Note that the assumption that

$$\Phi \in \mathbf{N} = \bigcup_{L \in \mathbb{N}} \bigcup_{(l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1}} \left(\prod_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right)$$

ensures that there exist $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$ which satisfy that

$$\Phi \in \left(\prod_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right). \quad (1.62)$$

Observe that (1.62), (1.55), and (1.56) imply that

$$\mathcal{L}(\Phi) = L, \quad \mathcal{I}(\Phi) = l_0 = \mathbb{D}_0(\Phi), \quad \text{and} \quad \mathcal{O}(\Phi) = l_L = \mathbb{D}_L(\Phi). \quad (1.63)$$

This shows that

$$\mathcal{D}(\Phi) = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1} = \mathbb{N}^{\mathcal{L}(\Phi)+1}. \quad (1.64)$$

Next note that (1.62), (1.56), and (1.58) ensure that for all $n \in \{1, 2, \dots, \mathcal{L}(\Phi)\}$ it holds that

$$\mathcal{W}_{n,\Phi} \in \mathbb{R}^{l_n \times l_{n-1}} = \mathbb{R}^{\mathbb{D}_n(\Phi) \times \mathbb{D}_{n-1}(\Phi)} \quad \text{and} \quad \mathcal{B}_{n,\Phi} \in \mathbb{R}^{l_n} = \mathbb{R}^{\mathbb{D}_n(\Phi)}. \quad (1.65)$$

The proof of Lemma 1.3.3 is thus complete. \square

1.3.2 Realizations of ANNs

Definition 1.3.4 (Realizations of ANNs). *Let $\Phi \in \mathbf{N}$ and let $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function (cf. Definition 1.3.1). Then we denote by*

$$\mathcal{R}_a^{\mathbf{N}}(\Phi): \mathbb{R}^{\mathcal{I}(\Phi)} \rightarrow \mathbb{R}^{\mathcal{O}(\Phi)} \quad (1.66)$$

the function which satisfies for all $x_0 \in \mathbb{R}^{\mathbb{D}_0(\Phi)}$, $x_1 \in \mathbb{R}^{\mathbb{D}_1(\Phi)}$, \dots , $x_{\mathcal{L}(\Phi)} \in \mathbb{R}^{\mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)}$ with

$$\forall k \in \{1, 2, \dots, \mathcal{L}(\Phi)\}: x_k = \mathfrak{M}_{a \mathbb{1}_{(0, \mathcal{L}(\Phi))}(k) + \text{id}_{\mathbb{R} \mathbb{1}_{\{\mathcal{L}(\Phi)\}}(k), \mathbb{D}_k(\Phi)}(\mathcal{W}_{k,\Phi} x_{k-1} + \mathcal{B}_{k,\Phi}) \quad (1.67)$$

that

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi))(x_0) = x_{\mathcal{L}(\Phi)} \quad (1.68)$$

*and we call $\mathcal{R}_a^{\mathbf{N}}(\Phi)$ the **realization** function of the ANN Φ with activation function a (cf. Definition 1.2.1).*

Remark 1.3.5 (Different uses of the term ANN in the literature). In Definition 1.3.2 above, we defined an ANN as a structured tuple of real numbers, or in other words, as a structured set of parameters. However, in the literature and colloquial usage, the term ANN sometimes also refers to a different mathematical object. Specifically,

for a given architecture and activation function, it may refer to the function that maps parameters and input to the output of the corresponding realization function. More formally, let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function, and consider the function

$$\mathcal{f}: \left(\prod_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right) \times \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L} \quad (1.69)$$

which satisfies for all $\Phi \in \left(\prod_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right)$, $x \in \mathbb{R}^{l_0}$ that

$$\mathcal{f}(\Phi, x) = \mathcal{R}_a^{\mathbf{N}}(\Phi)(x). \quad (1.70)$$

In this context, the function \mathcal{f} itself is sometimes referred to as an ANN.

Exercise 1.3.1. Let

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3)) \in (\mathbb{R}^{2 \times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}^1) \quad (1.71)$$

satisfy

$$W_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad W_2 = \begin{pmatrix} -1 & 2 \\ 3 & -4 \\ -5 & 6 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad (1.72)$$

$$W_3 = (-1 \ 1 \ -1), \quad \text{and} \quad B_3 = (-4). \quad (1.73)$$

Prove or disprove the following statement: It holds that

$$(\mathcal{R}_t^{\mathbf{N}}(\Phi))(-1) = 0 \quad (1.74)$$

(cf. Definitions 1.2.4 and 1.3.4).

1.3.3 On the connection to the vectorized description

Definition 1.3.6 (Transformation from the structured to the vectorized description of ANNs). We denote by $\mathcal{T}: \mathbf{N} \rightarrow \left(\bigcup_{d \in \mathbb{N}} \mathbb{R}^d \right)$ the function which satisfies for all

$\Phi \in \mathbf{N}$, $k \in \{1, 2, \dots, \mathcal{L}(\Phi)\}$, $d \in \mathbb{N}$, $\theta = (\theta_1, \theta_2, \dots, \theta_d) \in \mathbb{R}^d$ with $\mathcal{T}(\Phi) = \theta$ that

$$d = \mathcal{P}(\Phi), \quad \mathcal{B}_{k,\Phi} = \begin{pmatrix} \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_k l_{k-1} + 1} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_k l_{k-1} + 2} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_k l_{k-1} + 3} \\ \vdots \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_k l_{k-1} + l_k} \end{pmatrix}, \quad \text{and} \quad \mathcal{W}_{k,\Phi} = \begin{pmatrix} \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_{k-1}} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_{k-1} + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_{k-1} + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2l_{k-1}} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2l_{k-1} + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2l_{k-1} + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 3l_{k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + (l_k - 1)l_{k-1} + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + (l_k - 1)l_{k-1} + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_k l_{k-1}} \end{pmatrix} \quad (1.75)$$

(cf. Definition 1.3.1).

Example 1.3.7. Let $\Phi \in (\mathbb{R}^{3 \times 3} \times \mathbb{R}^3) \times (\mathbb{R}^{2 \times 3} \times \mathbb{R}^2)$ satisfy

$$\Phi = \left(\left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix} \right), \left(\begin{pmatrix} 13 & 14 & 15 \\ 16 & 17 & 18 \end{pmatrix}, \begin{pmatrix} 19 \\ 20 \end{pmatrix} \right) \right). \quad (1.76)$$

Then $\mathcal{T}(\Phi) = (1, 2, 3, \dots, 19, 20) \in \mathbb{R}^{20}$.

Lemma 1.3.8. Let $a, b \in \mathbb{N}$, $W = (W_{i,j})_{(i,j) \in \{1,2,\dots,a\} \times \{1,2,\dots,b\}} \in \mathbb{R}^{a \times b}$, $B = (B_1, B_2, \dots, B_a) \in \mathbb{R}^a$. Then

$$\begin{aligned} & \mathcal{T}(((W, B))) \\ &= (W_{1,1}, W_{1,2}, \dots, W_{1,b}, W_{2,1}, W_{2,2}, \dots, W_{2,b}, \dots, W_{a,1}, W_{a,2}, \dots, W_{a,b}, B_1, B_2, \dots, B_a) \end{aligned} \quad (1.77)$$

(cf. Definition 1.3.6).

Lemma 1.3.9. Let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$ and for every $k \in \{1, 2, \dots, L\}$ let $W_k = (W_{k,i,j})_{(i,j) \in \{1,2,\dots,l_k\} \times \{1,2,\dots,l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $B_k = (B_{k,1}, B_{k,2}, \dots, B_{k,l_k}) \in \mathbb{R}^{l_k}$.

Then

$$\begin{aligned}
 & \mathcal{T}\left(\left((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L)\right)\right) \\
 &= \left(W_{1,1,1}, W_{1,1,2}, \dots, W_{1,1,l_0}, \dots, W_{1,l_1,1}, W_{1,l_1,2}, \dots, W_{1,l_1,l_0}, B_{1,1}, B_{1,2}, \dots, B_{1,l_1}, \right. \\
 &\quad W_{2,1,1}, W_{2,1,2}, \dots, W_{2,1,l_1}, \dots, W_{2,l_2,1}, W_{2,l_2,2}, \dots, W_{2,l_2,l_1}, B_{2,1}, B_{2,2}, \dots, B_{2,l_2}, \\
 &\quad \dots, \\
 &\quad \left. W_{L,1,1}, W_{L,1,2}, \dots, W_{L,1,l_{L-1}}, \dots, W_{L,l_L,1}, W_{L,l_L,2}, \dots, W_{L,l_L,l_{L-1}}, B_{L,1}, B_{L,2}, \dots, B_{L,l_L}\right)
 \end{aligned} \tag{1.78}$$

(cf. Definition 1.3.6).

Exercise 1.3.2. Prove or disprove the following statement: The function \mathcal{T} is injective (cf. Definition 1.3.6).

Exercise 1.3.3. Prove or disprove the following statement: The function \mathcal{T} is surjective (cf. Definition 1.3.6).

Exercise 1.3.4. Prove or disprove the following statement: The function \mathcal{T} is bijective (cf. Definition 1.3.6).

Proposition 1.3.10. Let $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then

$$\mathcal{R}_a^{\mathbf{N}}(\Phi) = \begin{cases} \mathcal{N}_{\text{id}_{\mathbb{R}^{\mathcal{O}(\Phi)}}}^{\mathcal{T}(\Phi), \mathcal{I}(\Phi)} & : \mathcal{H}(\Phi) = 0 \\ \mathcal{N}_{\mathfrak{M}_{a, \mathbb{D}_1(\Phi)}, \mathfrak{M}_{a, \mathbb{D}_2(\Phi)}, \dots, \mathfrak{M}_{a, \mathbb{D}_{\mathcal{H}(\Phi)}(\Phi)}, \text{id}_{\mathbb{R}^{\mathcal{O}(\Phi)}}}^{\mathcal{T}(\Phi), \mathcal{I}(\Phi)} & : \mathcal{H}(\Phi) > 0 \end{cases} \tag{1.79}$$

(cf. Definitions 1.1.3, 1.2.1, 1.3.4, and 1.3.6).

Proof of Proposition 1.3.10. Throughout this proof, let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$ satisfy that

$$\mathcal{L}(\Phi) = L \quad \text{and} \quad \mathcal{D}(\Phi) = (l_0, l_1, \dots, l_L), \tag{1.80}$$

and let v_0, v_1, \dots, v_{L-1} satisfy for all $k \in \{0, 1, \dots, L-1\}$ that

$$v_k = \sum_{i=1}^k l_i (l_{i-1} + 1). \tag{1.81}$$

Note that (1.75) shows that for all $k \in \{1, 2, \dots, L\}$, $x \in \mathbb{R}^{l_{k-1}}$ it holds that

$$\mathcal{W}_{k, \Phi} x + \mathcal{B}_{k, \Phi} = \left(\mathcal{A}_{l_k, l_{k-1}}^{\mathcal{T}(\Phi), v_{k-1}}\right)(x) \tag{1.82}$$

(cf. Definitions 1.1.1 and 1.3.6). This demonstrates that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}$, \dots , $x_{L-1} \in \mathbb{R}^{l_{L-1}}$ with $\forall k \in \{1, 2, \dots, L-1\}$: $x_k = \mathfrak{M}_{a,l_k}(\mathcal{W}_{k,\Phi}x_{k-1} + \mathcal{B}_{k,\Phi})$ it holds that

$$x_{L-1} = \begin{cases} x_0 & : L = 1 \\ (\mathfrak{M}_{a,l_{L-1}} \circ \mathcal{A}_{l_{L-1},l_{L-2}}^{\mathcal{T}(\Phi),v_{L-2}} \circ \mathfrak{M}_{a,l_{L-2}} \circ \mathcal{A}_{l_{L-2},l_{L-3}}^{\mathcal{T}(\Phi),v_{L-3}} \circ \dots \circ \mathfrak{M}_{a,l_1} \circ \mathcal{A}_{l_1,l_0}^{\mathcal{T}(\Phi),0})(x_0) & : L > 1 \end{cases} \quad (1.83)$$

(cf. Definition 1.2.1). This, (1.82), (1.5), and (1.68) show that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}$, \dots , $x_L \in \mathbb{R}^{l_L}$ with $\forall k \in \{1, 2, \dots, L\}$: $x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k)+\text{id}_{\mathbb{R}^{\mathbb{1}_{\{L\}}}(k),l_k}}(\mathcal{W}_{k,\Phi}x_{k-1} + \mathcal{B}_{k,\Phi})$ it holds that

$$\begin{aligned} (\mathcal{R}_a^{\mathbb{N}}(\Phi))(x_0) &= x_L = \mathcal{W}_{L,\Phi}x_{L-1} + \mathcal{B}_{L,\Phi} = (\mathcal{A}_{l_L,l_{L-1}}^{\mathcal{T}(\Phi),v_{L-1}})(x_{L-1}) \\ &= \begin{cases} (\mathcal{N}_{\text{id}_{\mathbb{R}^{l_L}}^{\mathcal{T}(\Phi),l_0}})(x_0) & : L = 1 \\ (\mathcal{N}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\dots,\mathfrak{M}_{a,l_{L-1}},\text{id}_{\mathbb{R}^{l_L}}^{\mathcal{T}(\Phi),l_0}})(x_0) & : L > 1 \end{cases} \end{aligned} \quad (1.84)$$

(cf. Definitions 1.1.3 and 1.3.4). The proof of Proposition 1.3.10 is thus complete. \square

Chapter 2

ANN calculus

In this chapter we review certain operations that can be performed on the set of ANNs such as compositions (see Section 2.1), paralellizations (see Section 2.2), scalar multiplications (see Section 2.3), and sums (see Section 2.4) and thereby review an appropriate calculus for ANNs. The operations and the calculus for ANNs presented in this chapter will be used in Chapter 3 to establish certain ANN approximation results.

2.1 Compositions of ANNs

2.1.1 Compositions of ANNs

Definition 2.1.1 (Composition of ANNs). *We denote by*

$$(\cdot) \bullet (\cdot) : \{(\Phi, \Psi) \in \mathbf{N} \times \mathbf{N} : \mathcal{I}(\Phi) = \mathcal{O}(\Psi)\} \rightarrow \mathbf{N} \quad (2.1)$$

the function which satisfies for all $\Phi, \Psi \in \mathbf{N}$, $k \in \{1, 2, \dots, \mathcal{L}(\Phi) + \mathcal{L}(\Psi) - 1\}$ with $\mathcal{I}(\Phi) = \mathcal{O}(\Psi)$ that $\mathcal{L}(\Phi \bullet \Psi) = \mathcal{L}(\Phi) + \mathcal{L}(\Psi) - 1$ and

$$(\mathcal{W}_{k, \Phi \bullet \Psi}, \mathcal{B}_{k, \Phi \bullet \Psi}) = \begin{cases} (\mathcal{W}_{k, \Psi}, \mathcal{B}_{k, \Psi}) & : k < \mathcal{L}(\Psi) \\ (\mathcal{W}_{1, \Phi} \mathcal{W}_{\mathcal{L}(\Psi), \Psi}, \mathcal{W}_{1, \Phi} \mathcal{B}_{\mathcal{L}(\Psi), \Psi} + \mathcal{B}_{1, \Phi}) & : k = \mathcal{L}(\Psi) \\ (\mathcal{W}_{k - \mathcal{L}(\Psi) + 1, \Phi}, \mathcal{B}_{k - \mathcal{L}(\Psi) + 1, \Phi}) & : k > \mathcal{L}(\Psi) \end{cases} \quad (2.2)$$

(cf. Definition 1.3.1).

2.1.2 Elementary properties of compositions of ANNs

The next result shows that the set of neural network realisations is closed under composition. This property is central to the approximation theory of deep networks.

Proposition 2.1.2 (Properties of standard compositions of ANNs). *Let $\Phi, \Psi \in \mathbf{N}$ satisfy $\mathcal{I}(\Phi) = \mathcal{O}(\Psi)$ (cf. Definition 1.3.1). Then*

(i) *it holds that*

$$\mathcal{D}(\Phi \bullet \Psi) = (\mathbb{D}_0(\Psi), \mathbb{D}_1(\Psi), \dots, \mathbb{D}_{\mathcal{H}(\Psi)}(\Psi), \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \quad (2.3)$$

(ii) *it holds that*

$$[\mathcal{L}(\Phi \bullet \Psi) - 1] = [\mathcal{L}(\Phi) - 1] + [\mathcal{L}(\Psi) - 1], \quad (2.4)$$

(iii) *it holds that*

$$\mathcal{H}(\Phi \bullet \Psi) = \mathcal{H}(\Phi) + \mathcal{H}(\Psi), \quad (2.5)$$

(iv) *it holds that*

$$\begin{aligned} \mathcal{P}(\Phi \bullet \Psi) &= \mathcal{P}(\Phi) + \mathcal{P}(\Psi) + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{L}(\Psi)-1}(\Psi) + 1) \\ &\quad - \mathbb{D}_1(\Phi)(\mathbb{D}_0(\Phi) + 1) - \mathbb{D}_{\mathcal{L}(\Psi)}(\Psi)(\mathbb{D}_{\mathcal{L}(\Psi)-1}(\Psi) + 1) \\ &\leq \mathcal{P}(\Phi) + \mathcal{P}(\Psi) + \mathbb{D}_1(\Phi)\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi), \end{aligned} \quad (2.6)$$

and

(v) *it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \Psi) \in C(\mathbb{R}^{\mathcal{I}(\Psi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$ and*

$$\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \Psi) = [\mathcal{R}_a^{\mathbf{N}}(\Phi)] \circ [\mathcal{R}_a^{\mathbf{N}}(\Psi)] \quad (2.7)$$

(cf. Definitions 1.3.4 and 2.1.1).

Proof of Proposition 2.1.2. Throughout this proof let $L = \mathcal{L}(\Phi \bullet \Psi)$. Observe that the fact that $\mathcal{L}(\Phi \bullet \Psi) = \mathcal{L}(\Phi) + \mathcal{L}(\Psi) - 1$ and the fact that for all $\Theta \in \mathbf{N}$ it holds that $\mathcal{H}(\Theta) = \mathcal{L}(\Theta) - 1$ establish items (ii) and (iii). Note that item (iii) in Lemma 1.3.3 and (2.2) show that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$\mathcal{W}_{k, \Phi \bullet \Psi} \in \begin{cases} \mathbb{R}^{\mathbb{D}_k(\Psi) \times \mathbb{D}_{k-1}(\Psi)} & : k < \mathcal{L}(\Psi) \\ \mathbb{R}^{\mathbb{D}_1(\Phi) \times \mathbb{D}_{\mathcal{L}(\Psi)-1}(\Psi)} & : k = \mathcal{L}(\Psi) \\ \mathbb{R}^{\mathbb{D}_{k-\mathcal{L}(\Psi)+1}(\Phi) \times \mathbb{D}_{k-\mathcal{L}(\Psi)}(\Phi)} & : k > \mathcal{L}(\Psi). \end{cases} \quad (2.8)$$

This, item (iii) in Lemma 1.3.3, and the fact that $\mathcal{H}(\Psi) = \mathcal{L}(\Psi) - 1$ ensure that for all

$k \in \{0, 1, \dots, L\}$ it holds that

$$\mathbb{D}_k(\Phi \bullet \Psi) = \begin{cases} \mathbb{D}_k(\Psi) & : k \leq \mathcal{H}(\Psi) \\ \mathbb{D}_{k-\mathcal{L}(\Psi)+1}(\Phi) & : k > \mathcal{H}(\Psi). \end{cases} \quad (2.9)$$

This establishes item (i). Observe that (2.9) implies that

$$\begin{aligned} \mathcal{P}(\Phi_1 \bullet \Phi_2) &= \sum_{j=1}^L \mathbb{D}_j(\Phi \bullet \Psi)(\mathbb{D}_{j-1}(\Phi \bullet \Psi) + 1) \\ &= \left[\sum_{j=1}^{\mathcal{H}(\Psi)} \mathbb{D}_j(\Psi)(\mathbb{D}_{j-1}(\Psi) + 1) \right] + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi) + 1) \\ &\quad + \left[\sum_{j=\mathcal{L}(\Psi)+1}^L \mathbb{D}_{j-\mathcal{L}(\Psi)+1}(\Phi)(\mathbb{D}_{j-\mathcal{L}(\Psi)}(\Phi) + 1) \right] \\ &= \left[\sum_{j=1}^{\mathcal{L}(\Psi)-1} \mathbb{D}_j(\Psi)(\mathbb{D}_{j-1}(\Psi) + 1) \right] + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi) + 1) \\ &\quad + \left[\sum_{j=2}^{\mathcal{L}(\Phi)} \mathbb{D}_j(\Phi)(\mathbb{D}_{j-1}(\Phi) + 1) \right] \\ &= [\mathcal{P}(\Psi) - \mathbb{D}_{\mathcal{L}(\Psi)}(\Psi)(\mathbb{D}_{\mathcal{L}(\Psi)-1}(\Psi) + 1)] + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi) + 1) \\ &\quad + [\mathcal{P}(\Phi) - \mathbb{D}_1(\Phi)(\mathbb{D}_0(\Phi) + 1)]. \end{aligned} \quad (2.10)$$

This proves item (iv). The proof of item (v) mainly uses the recursive definition of realization function and the associativity of function composition, and is left as an exercise. The proof of Proposition 2.1.2 is thus complete. \square

2.1.3 Associativity of compositions of ANNs

The next result shows an important property of composition of ANNs, i.e. that it is associative. Its proof follows directly from Definition 2.1.1 and is left as an exercise.

Lemma 2.1.3. *Let $\Phi_1, \Phi_2, \Phi_3 \in \mathbf{N}$ satisfy $\mathcal{I}(\Phi_1) = \mathcal{O}(\Phi_2)$ and $\mathcal{I}(\Phi_2) = \mathcal{O}(\Phi_3)$ (cf. Definition 1.3.1). Then*

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \quad (2.11)$$

(cf. Definition 2.1.1).

2.1.4 Powers of ANNs

Definition 2.1.4 (Powers of ANNs). We denote by $(\cdot)^{\bullet n}: \{\Phi \in \mathbf{N}: \mathcal{I}(\Phi) = \mathcal{O}(\Phi)\} \rightarrow \mathbf{N}$, $n \in \mathbb{N}_0$, the functions which satisfy for all $n \in \mathbb{N}_0$, $\Phi \in \mathbf{N}$ with $\mathcal{I}(\Phi) = \mathcal{O}(\Phi)$ that

$$\Phi^{\bullet n} = \begin{cases} (\mathbb{I}_{\mathcal{O}(\Phi)}, (0, 0, \dots, 0)) \in \mathbb{R}^{\mathcal{O}(\Phi) \times \mathcal{O}(\Phi)} \times \mathbb{R}^{\mathcal{O}(\Phi)} & : n = 0 \\ \Phi \bullet (\Phi^{\bullet(n-1)}) & : n \in \mathbb{N} \end{cases} \quad (2.12)$$

(cf. Definitions 1.3.1, 2.1.1, and 3.2.1).

Lemma 2.1.5 (Number of hidden layers of powers of ANNs). Let $n \in \mathbb{N}_0$, $\Phi \in \mathbf{N}$ satisfy $\mathcal{I}(\Phi) = \mathcal{O}(\Phi)$ (cf. Definition 1.3.1). Then

$$\mathcal{H}(\Phi^{\bullet n}) = n\mathcal{H}(\Phi) \quad (2.13)$$

(cf. Definition 2.1.4).

Proof of Lemma 2.1.5. Note that Proposition 2.1.2, (2.12), and induction establish (2.13). The proof of Lemma 2.1.5 is thus complete. \square

2.2 Parallelizations of ANNs

2.2.1 Parallelizations of ANNs with the same length

Definition 2.2.1 (Parallelization of ANNs). Let $n \in \mathbb{N}$. Then we denote by

$$\mathbf{P}_n: \{\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n: \mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)\} \rightarrow \mathbf{N} \quad (2.14)$$

the function which satisfies for all $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n$, $k \in \{1, 2, \dots, \mathcal{L}(\Phi_1)\}$ with

$\mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)$ that

$$\mathcal{L}(\mathbf{P}_n(\Phi)) = \mathcal{L}(\Phi_1), \quad \mathcal{W}_{k,\mathbf{P}_n(\Phi)} = \begin{pmatrix} \mathcal{W}_{k,\Phi_1} & 0 & 0 & \dots & 0 \\ 0 & \mathcal{W}_{k,\Phi_2} & 0 & \dots & 0 \\ 0 & 0 & \mathcal{W}_{k,\Phi_3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathcal{W}_{k,\Phi_n} \end{pmatrix},$$

$$\text{and } \mathcal{B}_{k,\mathbf{P}_n(\Phi)} = \begin{pmatrix} \mathcal{B}_{k,\Phi_1} \\ \mathcal{B}_{k,\Phi_2} \\ \vdots \\ \mathcal{B}_{k,\Phi_n} \end{pmatrix} \quad (2.15)$$

(cf. Definition 1.3.1).

Lemma 2.2.2 (Architectures of parallelizations of ANNs). *Let $n, L \in \mathbb{N}$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbb{N}^n$ satisfy $L = \mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)$ (cf. Definition 1.3.1). Then*

(i) *it holds that*

$$\mathbf{P}_n(\Phi) \in \left(\prod_{k=1}^L \left(\mathbb{R}^{(\sum_{j=1}^n \mathbb{D}_k(\Phi_j)) \times (\sum_{j=1}^n \mathbb{D}_{k-1}(\Phi_j))} \times \mathbb{R}^{(\sum_{j=1}^n \mathbb{D}_k(\Phi_j))} \right) \right), \quad (2.16)$$

(ii) *it holds for all $k \in \mathbb{N}_0$ that*

$$\mathbb{D}_k(\mathbf{P}_n(\Phi)) = \mathbb{D}_k(\Phi_1) + \mathbb{D}_k(\Phi_2) + \dots + \mathbb{D}_k(\Phi_n), \quad (2.17)$$

and

(iii) *it holds that*

$$\mathcal{D}(\mathbf{P}_n(\Phi)) = \mathcal{D}(\Phi_1) + \mathcal{D}(\Phi_2) + \dots + \mathcal{D}(\Phi_n) \quad (2.18)$$

(cf. Definition 2.2.1).

Proof of Lemma 2.2.2. Note that item (iii) in Lemma 1.3.3 and (2.15) imply that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$\mathcal{W}_{k,\mathbf{P}_n(\Phi)} \in \mathbb{R}^{(\sum_{j=1}^n \mathbb{D}_k(\Phi_j)) \times (\sum_{j=1}^n \mathbb{D}_{k-1}(\Phi_j))} \quad \text{and} \quad \mathcal{B}_{k,\mathbf{P}_n(\Phi)} \in \mathbb{R}^{(\sum_{j=1}^n \mathbb{D}_k(\Phi_j))} \quad (2.19)$$

(cf. Definition 2.2.1). Item (iii) in Lemma 1.3.3 therefore establishes items (i) and (ii). Observe that item (ii) implies item (iii). The proof of Lemma 2.2.2 is thus complete. \square

Proposition 2.2.3 (Realizations of parallelizations of ANNs). *Let $a \in C(\mathbb{R}, \mathbb{R})$, $n \in \mathbb{N}$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n$ satisfy $\mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)$ (cf. Definition 1.3.1). Then*

(i) *it holds that*

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\Phi)) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.20)$$

and

(ii) *it holds for all $x_1 \in \mathbb{R}^{\mathcal{I}(\Phi_1)}$, $x_2 \in \mathbb{R}^{\mathcal{I}(\Phi_2)}$, \dots , $x_n \in \mathbb{R}^{\mathcal{I}(\Phi_n)}$ that*

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\Phi)))(x_1, x_2, \dots, x_n) \\ &= ((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n)) \in \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]} \end{aligned} \quad (2.21)$$

(cf. Definitions 1.3.4 and 2.2.1).

Proof of Proposition 2.2.3. Throughout this proof let $L = \mathcal{L}(\Phi_1)$. Note that item (ii) in Lemma 2.2.2 and item (ii) in Lemma 1.3.3 imply that

$$\mathcal{I}(\mathbf{P}_n(\Phi)) = \mathbb{D}_0(\mathbf{P}_n(\Phi)) = \sum_{j=1}^n \mathbb{D}_0(\Phi_j) = \sum_{j=1}^n \mathcal{I}(\Phi_j). \quad (2.22)$$

Next observe that item (ii) in Lemma 2.2.2 and item (ii) in Lemma 1.3.3 ensure that

$$\mathcal{O}(\mathbf{P}_n(\Phi)) = \mathbb{D}_{\mathcal{L}(\mathbf{P}_n(\Phi))}(\mathbf{P}_n(\Phi)) = \sum_{j=1}^n \mathbb{D}_{\mathcal{L}(\Phi_j)}(\Phi_j) = \sum_{j=1}^n \mathcal{O}(\Phi_j). \quad (2.23)$$

Note that (2.15) and item (ii) in Lemma 2.2.2 show that for all $k \in \{1, 2, \dots, L\}$, $x_1 \in \mathbb{R}^{\mathbb{D}_k(\Phi_1)}$, $x_2 \in \mathbb{R}^{\mathbb{D}_k(\Phi_2)}$, \dots , $x_n \in \mathbb{R}^{\mathbb{D}_k(\Phi_n)}$, $y \in \mathbb{R}^{[\sum_{j=1}^n \mathbb{D}_k(\Phi_j)]}$ with $y = (x_1, x_2, \dots, x_n)$ it holds

that

$$\begin{aligned}
 & \mathfrak{M}_{a, \mathbb{D}_k(\mathbf{P}_n(\Phi))}(\mathcal{W}_{k, \mathbf{P}_n(\Phi)} y + \mathcal{B}_{k, \mathbf{P}_n(\Phi)}) \\
 &= \mathfrak{M}_{a, \mathbb{D}_k(\mathbf{P}_n(\Phi))} \left(\begin{pmatrix} \mathcal{W}_{k, \Phi_1} & 0 & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{k, \Phi_2} & 0 & \cdots & 0 \\ 0 & 0 & \mathcal{W}_{k, \Phi_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathcal{W}_{k, \Phi_n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \mathcal{B}_{k, \Phi_1} \\ \mathcal{B}_{k, \Phi_2} \\ \mathcal{B}_{k, \Phi_3} \\ \vdots \\ \mathcal{B}_{k, \Phi_n} \end{pmatrix} \right) \\
 &= \mathfrak{M}_{a, \mathbb{D}_k(\mathbf{P}_n(\Phi))} \left(\begin{pmatrix} \mathcal{W}_{k, \Phi_1} x_1 + \mathcal{B}_{k, \Phi_1} \\ \mathcal{W}_{k, \Phi_2} x_2 + \mathcal{B}_{k, \Phi_2} \\ \mathcal{W}_{k, \Phi_3} x_3 + \mathcal{B}_{k, \Phi_3} \\ \vdots \\ \mathcal{W}_{k, \Phi_n} x_n + \mathcal{B}_{k, \Phi_n} \end{pmatrix} \right) = \begin{pmatrix} \mathfrak{M}_{a, \mathbb{D}_k(\Phi_1)}(\mathcal{W}_{k, \Phi_1} x_1 + \mathcal{B}_{k, \Phi_1}) \\ \mathfrak{M}_{a, \mathbb{D}_k(\Phi_2)}(\mathcal{W}_{k, \Phi_2} x_2 + \mathcal{B}_{k, \Phi_2}) \\ \mathfrak{M}_{a, \mathbb{D}_k(\Phi_3)}(\mathcal{W}_{k, \Phi_3} x_3 + \mathcal{B}_{k, \Phi_3}) \\ \vdots \\ \mathfrak{M}_{a, \mathbb{D}_k(\Phi_n)}(\mathcal{W}_{k, \Phi_n} x_n + \mathcal{B}_{k, \Phi_n}) \end{pmatrix}. \tag{2.24}
 \end{aligned}$$

Induction and (1.68) hence demonstrate that for all $x_1 \in \mathbb{R}^{\mathcal{I}(\Phi_1)}$, $x_2 \in \mathbb{R}^{\mathcal{I}(\Phi_2)}$, \dots , $x_n \in \mathbb{R}^{\mathcal{I}(\Phi_n)}$ that

$$\begin{aligned}
 & (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\Phi)))(x_1, x_2, \dots, x_n) \\
 &= ((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n)) \in \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]} \tag{2.25}
 \end{aligned}$$

This establishes item (ii). The proof of Proposition 2.2.3 is thus complete. \square

Proposition 2.2.4 (Upper bounds for the numbers of parameters of parallelizations of ANNs). *Let $n, L \in \mathbb{N}$, $\Phi_1, \Phi_2, \dots, \Phi_n \in \mathbf{N}$ satisfy $L = \mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)$ (cf. Definition 1.3.1). Then*

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \dots, \Phi_n)) \leq \frac{1}{2} [\sum_{j=1}^n \mathcal{P}(\Phi_j)]^2 \tag{2.26}$$

(cf. Definition 2.2.1).

Proof of Proposition 2.2.4. Throughout this proof, for every $j \in \{1, 2, \dots, n\}$, $k \in \{0, 1,$

$\dots, L\}$ let $l_{j,k} = \mathbb{D}_k(\Phi_j)$. Observe that item (ii) in Lemma 2.2.2 demonstrates that

$$\begin{aligned}
 \mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \dots, \Phi_n)) &= \sum_{k=1}^L \left[\sum_{i=1}^n l_{i,k} \right] \left[\left(\sum_{i=1}^n l_{i,k-1} \right) + 1 \right] \\
 &= \sum_{k=1}^L \left[\sum_{i=1}^n l_{i,k} \right] \left[\left(\sum_{j=1}^n l_{j,k-1} \right) + 1 \right] \\
 &\leq \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^L l_{i,k} (l_{j,k-1} + 1) \leq \sum_{i=1}^n \sum_{j=1}^n \sum_{k,\ell=1}^L l_{i,k} (l_{j,\ell-1} + 1) \\
 &= \sum_{i=1}^n \sum_{j=1}^n \left[\sum_{k=1}^L l_{i,k} \right] \left[\sum_{\ell=1}^L (l_{j,\ell-1} + 1) \right] \\
 &\leq \sum_{i=1}^n \sum_{j=1}^n \left[\sum_{k=1}^L \frac{1}{2} l_{i,k} (l_{i,k-1} + 1) \right] \left[\sum_{\ell=1}^L l_{j,\ell} (l_{j,\ell-1} + 1) \right] \\
 &= \sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} \mathcal{P}(\Phi_i) \mathcal{P}(\Phi_j) = \frac{1}{2} \left[\sum_{i=1}^n \mathcal{P}(\Phi_i) \right]^2.
 \end{aligned} \tag{2.27}$$

The proof of Proposition 2.2.4 is thus complete. \square

Corollary 2.2.5 (Lower and upper bounds for the numbers of parameters of parallelizations of ANNs). *Let $n \in \mathbb{N}$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n$ satisfy $\mathcal{D}(\Phi_1) = \mathcal{D}(\Phi_2) = \dots = \mathcal{D}(\Phi_n)$ (cf. Definition 1.3.1). Then*

$$\left[\frac{n^2}{2} \right] \mathcal{P}(\Phi_1) \leq \left[\frac{n^2+n}{2} \right] \mathcal{P}(\Phi_1) \leq \mathcal{P}(\mathbf{P}_n(\Phi)) \leq n^2 \mathcal{P}(\Phi_1) \leq \frac{1}{2} \left[\sum_{i=1}^n \mathcal{P}(\Phi_i) \right]^2 \tag{2.28}$$

(cf. Definition 2.2.1).

Proof of Corollary 2.2.5. Throughout this proof, let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$ satisfy

$$\mathcal{D}(\Phi_1) = (l_0, l_1, \dots, l_L). \tag{2.29}$$

Note that (2.29) and the assumption that $\mathcal{D}(\Phi_1) = \mathcal{D}(\Phi_2) = \dots = \mathcal{D}(\Phi_n)$ imply that for all $j \in \{1, 2, \dots, n\}$ it holds that

$$\mathcal{D}(\Phi_j) = (l_0, l_1, \dots, l_L). \tag{2.30}$$

Combining this with item (iii) in Lemma 2.2.2 demonstrates that

$$\mathcal{P}(\mathbf{P}_n(\Phi)) = \sum_{j=1}^L (nl_j) ((nl_{j-1}) + 1). \tag{2.31}$$

Hence, we obtain that

$$\mathcal{P}(\mathbf{P}_n(\Phi)) \leq \sum_{j=1}^L (nl_j)((nl_{j-1}) + n) = n^2 \left[\sum_{j=1}^L l_j(l_{j-1} + 1) \right] = n^2 \mathcal{P}(\Phi_1). \quad (2.32)$$

Next observe that the assumption that $\mathcal{D}(\Phi_1) = \mathcal{D}(\Phi_2) = \dots = \mathcal{D}(\Phi_n)$ and the fact that $\mathcal{P}(\Phi_1) \geq l_1(l_0 + 1) \geq 2$ ensure that

$$n^2 \mathcal{P}(\Phi_1) \leq \frac{n^2}{2} [\mathcal{P}(\Phi_1)]^2 = \frac{1}{2} [n \mathcal{P}(\Phi_1)]^2 = \frac{1}{2} \left[\sum_{i=1}^n \mathcal{P}(\Phi_1) \right]^2 = \frac{1}{2} \left[\sum_{i=1}^n \mathcal{P}(\Phi_i) \right]^2. \quad (2.33)$$

Furthermore, note that (2.31) and the fact that for all $a, b \in \mathbb{N}$ it holds that

$$2(ab + 1) = ab + 1 + (a - 1)(b - 1) + a + b \geq ab + a + b + 1 = (a + 1)(b + 1) \quad (2.34)$$

show that

$$\begin{aligned} \mathcal{P}(\mathbf{P}_n(\Phi)) &\geq \frac{1}{2} \left[\sum_{j=1}^L (nl_j)(n + 1)(l_{j-1} + 1) \right] \\ &= \frac{n(n+1)}{2} \left[\sum_{j=1}^L l_j(l_{j-1} + 1) \right] = \left[\frac{n^2+n}{2} \right] \mathcal{P}(\Phi_1). \end{aligned} \quad (2.35)$$

This, (2.32), and (2.33) establish (2.28). The proof of Corollary 2.2.5 is thus complete. \square

Exercise 2.2.1. Prove or disprove the following statement: For every $n \in \mathbb{N}$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n$ with $\mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)$ it holds that

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \dots, \Phi_n)) \leq n \left[\sum_{i=1}^n \mathcal{P}(\Phi_i) \right]. \quad (2.36)$$

Exercise 2.2.2. Prove or disprove the following statement: For every $n \in \mathbb{N}$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n$ with $\mathcal{P}(\Phi_1) = \mathcal{P}(\Phi_2) = \dots = \mathcal{P}(\Phi_n)$ it holds that

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \dots, \Phi_n)) \leq n^2 \mathcal{P}(\Phi_1). \quad (2.37)$$

2.2.2 Representations of the identities with ReLU activation functions

Definition 2.2.6 (ReLU identity ANNs). *We denote by $\mathfrak{J}_d \in \mathbf{N}$, $d \in \mathbb{N}$, the ANNs which satisfy for all $d \in \mathbb{N}$ that*

$$\mathfrak{J}_1 = \left(\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right), \left((1 \ -1), 0 \right) \right) \in ((\mathbb{R}^{2 \times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{1 \times 2} \times \mathbb{R}^1)) \quad (2.38)$$

and

$$\mathfrak{J}_d = \mathbf{P}_d(\mathfrak{J}_1, \mathfrak{J}_1, \dots, \mathfrak{J}_1) \quad (2.39)$$

(cf. Definitions 1.3.1 and 2.2.1).

Lemma 2.2.7 (Properties of ReLU identity ANNs). *Let $d \in \mathbb{N}$. Then*

(i) *it holds that*

$$\mathcal{D}(\mathfrak{J}_d) = (d, 2d, d) \in \mathbb{N}^3 \quad (2.40)$$

and

(ii) *it holds that*

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_d) = \text{id}_{\mathbb{R}^d} \quad (2.41)$$

(cf. Definitions 1.3.1, 1.3.4, and 2.2.6).

Proof of Lemma 2.2.7. Throughout this proof, let $L = 2$, $l_0 = 1$, $l_1 = 2$, $l_2 = 1$. Note that (2.38) ensures that

$$\mathcal{D}(\mathfrak{J}_1) = (1, 2, 1) = (l_0, l_1, l_2). \quad (2.42)$$

This, (2.39), and Proposition 2.2.4 prove that

$$\mathcal{D}(\mathfrak{J}_d) = (d, 2d, d) \in \mathbb{N}^3. \quad (2.43)$$

This establishes item (i). Next note that (2.38) assures that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_1))(x) = \mathfrak{r}(x) - \mathfrak{r}(-x) = \max\{x, 0\} - \max\{-x, 0\} = x. \quad (2.44)$$

Combining this and Proposition 2.2.3 demonstrates that for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ it holds that $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_d) \in C(\mathbb{R}^d, \mathbb{R}^d)$ and

$$\begin{aligned} (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_d))(x) &= (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{P}_d(\mathfrak{J}_1, \mathfrak{J}_1, \dots, \mathfrak{J}_1)))(x_1, x_2, \dots, x_d) \\ &= ((\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_1))(x_1), (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_1))(x_2), \dots, (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_1))(x_d)) \\ &= (x_1, x_2, \dots, x_d) = x \end{aligned} \quad (2.45)$$

(cf. Definition 2.2.1). This establishes item (ii). The proof of Lemma 2.2.7 is thus complete. \square

2.2.3 Extensions of ANNs

Definition 2.2.8 (Extensions of ANNs). *Let $L \in \mathbb{N}$, $\Psi \in \mathbf{N}$ satisfy $\mathcal{I}(\Psi) = \mathcal{O}(\Psi)$. Then we denote by*

$$\mathcal{E}_{L,\Psi}: \{\Phi \in \mathbf{N}: (\mathcal{L}(\Phi) \leq L \text{ and } \mathcal{O}(\Phi) = \mathcal{I}(\Psi))\} \rightarrow \mathbf{N} \quad (2.46)$$

the function which satisfies for all $\Phi \in \mathbf{N}$ with $\mathcal{L}(\Phi) \leq L$ and $\mathcal{O}(\Phi) = \mathcal{I}(\Psi)$ that

$$\mathcal{E}_{L,\Psi}(\Phi) = (\Psi^{\bullet(L-\mathcal{L}(\Phi))}) \bullet \Phi \quad (2.47)$$

(cf. Definitions 1.3.1, 2.1.1, and 2.1.4).

Lemma 2.2.9 (Length of extensions of ANNs). *Let $d, \mathbf{i} \in \mathbb{N}$, $\Psi \in \mathbf{N}$ satisfy $\mathcal{D}(\Psi) = (d, \mathbf{i}, d)$ (cf. Definition 1.3.1). Then*

(i) it holds for all $n \in \mathbb{N}_0$ that $\mathcal{H}(\Psi^{\bullet n}) = n$, $\mathcal{L}(\Psi^{\bullet n}) = n + 1$, $\mathcal{D}(\Psi^{\bullet n}) \in \mathbb{N}^{n+2}$, and

$$\mathcal{D}(\Psi^{\bullet n}) = \begin{cases} (d, d) & : n = 0 \\ (d, \mathbf{i}, \mathbf{i}, \dots, \mathbf{i}, d) & : n \in \mathbb{N} \end{cases} \quad (2.48)$$

and

(ii) it holds for all $\Phi \in \mathbf{N}$, $L \in \mathbb{N} \cap [\mathcal{L}(\Phi), \infty)$ with $\mathcal{O}(\Phi) = d$ that

$$\mathcal{L}(\mathcal{E}_{L,\Psi}(\Phi)) = L \quad (2.49)$$

(cf. Definitions 2.1.4 and 2.2.8).

Proof of Lemma 2.2.9. Throughout this proof, let $\Phi \in \mathbf{N}$ satisfy $\mathcal{O}(\Phi) = d$. Observe that Lemma 2.1.5 and the fact that $\mathcal{H}(\Psi) = 1$ imply that for all $n \in \mathbb{N}_0$ it holds that

$$\mathcal{H}(\Psi^{\bullet n}) = n\mathcal{H}(\Psi) = n \quad (2.50)$$

(cf. Definition 2.1.4). Combining this with (1.55) and Lemma 1.3.3 proves that

$$\mathcal{H}(\Psi^{\bullet n}) = n, \quad \mathcal{L}(\Psi^{\bullet n}) = n + 1, \quad \text{and} \quad \mathcal{D}(\Psi^{\bullet n}) \in \mathbb{N}^{n+2}. \quad (2.51)$$

Next we claim that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{N}^{n+2} \ni \mathcal{D}(\Psi^{\bullet n}) = \begin{cases} (d, d) & : n = 0 \\ (d, \mathbf{i}, \mathbf{i}, \dots, \mathbf{i}, d) & : n \in \mathbb{N}. \end{cases} \quad (2.52)$$

We now prove (2.52) by induction on $n \in \mathbb{N}_0$. Note that the fact that

$$\Psi^{\bullet 0} = (\mathbf{I}_d, 0) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \quad (2.53)$$

establishes (2.52) in the base case $n = 0$ (cf. Definition 3.2.1). For the induction step assume that there exists $n \in \mathbb{N}_0$ which satisfies

$$\mathbb{N}^{n+2} \ni \mathcal{D}(\Psi^{\bullet n}) = \begin{cases} (d, d) & : n = 0 \\ (d, \mathbf{i}, \mathbf{i}, \dots, \mathbf{i}, d) & : n \in \mathbb{N}. \end{cases} \quad (2.54)$$

Note that (2.54), (2.12), (2.51), item (i) in Proposition 2.1.2, and the fact that $\mathcal{D}(\Psi) = (d, \mathbf{i}, d) \in \mathbb{N}^3$ imply that

$$\mathcal{D}(\Psi^{\bullet(n+1)}) = \mathcal{D}(\Psi \bullet (\Psi^{\bullet n})) = (d, \mathbf{i}, \mathbf{i}, \dots, \mathbf{i}, d) \in \mathbb{N}^{n+3} \quad (2.55)$$

(cf. Definition 2.1.1). Induction therefore proves (2.52). This and (2.51) establish item (i). Observe that (2.47), item (iii) in Proposition 2.1.2, (2.50), and the fact that $\mathcal{H}(\Phi) = \mathcal{L}(\Phi) - 1$ demonstrate that for all $L \in \mathbb{N} \cap [\mathcal{L}(\Phi), \infty)$ it holds that

$$\begin{aligned} \mathcal{H}(\mathcal{E}_{L, \Psi}(\Phi)) &= \mathcal{H}((\Psi^{\bullet(L-\mathcal{L}(\Phi))}) \bullet \Phi) = \mathcal{H}(\Psi^{\bullet(L-\mathcal{L}(\Phi))}) + \mathcal{H}(\Phi) \\ &= (L - \mathcal{L}(\Phi)) + \mathcal{H}(\Phi) = L - 1. \end{aligned} \quad (2.56)$$

The fact that $\mathcal{H}(\mathcal{E}_{L, \Psi}(\Phi)) = \mathcal{L}(\mathcal{E}_{L, \Psi}(\Phi)) - 1$ hence establishes that

$$\mathcal{L}(\mathcal{E}_{L, \Psi}(\Phi)) = \mathcal{H}(\mathcal{E}_{L, \Psi}(\Phi)) + 1 = L. \quad (2.57)$$

This establishes item (ii). The proof of Lemma 2.2.9 is thus complete. \square

Lemma 2.2.10 (Realizations of extensions of ANNs). *Let $a \in C(\mathbb{R}, \mathbb{R})$, $\mathbb{I} \in \mathbf{N}$ satisfy $\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}) = \text{id}_{\mathbb{R}^{\mathcal{I}(\mathbb{I})}}$ (cf. Definitions 1.3.1 and 1.3.4). Then*

(i) *it holds for all $n \in \mathbb{N}_0$ that*

$$\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}) = \text{id}_{\mathbb{R}^{\mathcal{I}(\mathbb{I})}} \quad (2.58)$$

and

(ii) *it holds for all $\Phi \in \mathbf{N}$, $L \in \mathbb{N} \cap [\mathcal{L}(\Phi), \infty)$ with $\mathcal{O}(\Phi) = \mathcal{I}(\mathbb{I})$ that*

$$\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L, \mathbb{I}}(\Phi)) = \mathcal{R}_a^{\mathbf{N}}(\Phi) \quad (2.59)$$

(cf. Definitions 2.1.4 and 2.2.8).

Proof of Lemma 2.2.10. Throughout this proof, let $\Phi \in \mathbf{N}$, $L, d \in \mathbb{N}$ satisfy $\mathcal{L}(\Phi) \leq L$ and $\mathcal{I}(\mathbb{I}) = \mathcal{O}(\Phi) = d$. We claim that for all $n \in \mathbb{N}_0$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}) \in C(\mathbb{R}^d, \mathbb{R}^d) \quad \text{and} \quad \forall x \in \mathbb{R}^d: (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}))(x) = x. \quad (2.60)$$

We now prove (2.60) by induction on $n \in \mathbb{N}_0$. Note that (2.12) and the fact that $\mathcal{O}(\mathbb{I}) = d$ demonstrate that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet 0}) \in C(\mathbb{R}^d, \mathbb{R}^d)$ and $\forall x \in \mathbb{R}^d: (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet 0}))(x) = x$. This establishes (2.60) in the base case $n = 0$. For the induction step observe that for all $n \in \mathbb{N}_0$ with $\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}) \in C(\mathbb{R}^d, \mathbb{R}^d)$ and $\forall x \in \mathbb{R}^d: (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}))(x) = x$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(n+1)}) = \mathcal{R}_a^{\mathbf{N}}(\mathbb{I} \bullet (\mathbb{I}^{\bullet n})) = (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I})) \circ (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n})) \in C(\mathbb{R}^d, \mathbb{R}^d) \quad (2.61)$$

and

$$\begin{aligned} \forall x \in \mathbb{R}^d: (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(n+1)}))(x) &= ([\mathcal{R}_a^{\mathbf{N}}(\mathbb{I})] \circ [\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n})])(x) \\ &= (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}))((\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}))(x)) = (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}))(x) = x. \end{aligned} \quad (2.62)$$

Induction therefore proves (2.60). This establishes item (i). Note (2.47), item (v) in Proposition 2.1.2, item (i), and the fact that $\mathcal{I}(\mathbb{I}) = \mathcal{O}(\Phi)$ ensure that

$$\begin{aligned} \mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L, \mathbb{I}}(\Phi)) &= \mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(L-\mathcal{L}(\Phi))} \bullet \Phi) \\ &\in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\mathbb{I})}) = C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{I}(\mathbb{I})}) = C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)}) \end{aligned} \quad (2.63)$$

and

$$\begin{aligned} \forall x \in \mathbb{R}^{\mathcal{I}(\Phi)}: (\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L, \mathbb{I}}(\Phi)))(x) &= (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(L-\mathcal{L}(\Phi))})((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x))) \\ &= (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x). \end{aligned} \quad (2.64)$$

This establishes item (ii). The proof of Lemma 2.2.10 is thus complete. \square

Lemma 2.2.11 (Architectures of extensions of ANNs). *Let $d, \mathbf{i}, L, \mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \dots, l_{L-1} \in \mathbb{N}$, $\Phi, \Psi \in \mathbf{N}$ satisfy*

$$\mathfrak{L} \geq L, \quad \mathcal{D}(\Phi) = (l_0, l_1, \dots, l_{L-1}, d), \quad \text{and} \quad \mathcal{D}(\Psi) = (d, \mathbf{i}, d) \quad (2.65)$$

(cf. Definition 1.3.1). Then $\mathcal{D}(\mathcal{E}_{\mathfrak{L}, \Psi}(\Phi)) \in \mathbb{N}^{\mathfrak{L}+1}$ and

$$\mathcal{D}(\mathcal{E}_{\mathfrak{L}, \Psi}(\Phi)) = \begin{cases} (l_0, l_1, \dots, l_{L-1}, d) & : \mathfrak{L} = L \\ (l_0, l_1, \dots, l_{L-1}, \mathbf{i}, \mathbf{i}, \dots, \mathbf{i}, d) & : \mathfrak{L} > L \end{cases} \quad (2.66)$$

(cf. Definition 2.2.8).

Proof of Lemma 2.2.11. Observe that item (i) in Lemma 2.2.9 shows that

$$\mathcal{H}(\Psi^{\bullet(\mathfrak{L}-L)}) = \mathfrak{L} - L, \quad \mathcal{D}(\Psi^{\bullet(\mathfrak{L}-L)}) \in \mathbb{N}^{\mathfrak{L}-L+2}, \quad (2.67)$$

$$\text{and} \quad \mathcal{D}(\Psi^{\bullet(\mathfrak{L}-L)}) = \begin{cases} (d, d) & : \mathfrak{L} = L \\ (d, i, i, \dots, i, d) & : \mathfrak{L} > L \end{cases} \quad (2.68)$$

(cf. Definition 2.1.4). Combining this with Proposition 2.1.2 ensures that

$$\mathcal{H}((\Psi^{\bullet(\mathfrak{L}-L)}) \bullet \Phi) = \mathcal{H}(\Psi^{\bullet(\mathfrak{L}-L)}) + \mathcal{H}(\Phi) = (\mathfrak{L} - L) + L - 1 = \mathfrak{L} - 1, \quad (2.69)$$

$$\mathcal{D}((\Psi^{\bullet(\mathfrak{L}-L)}) \bullet \Phi) \in \mathbb{N}^{\mathfrak{L}+1}, \quad (2.70)$$

$$\text{and} \quad \mathcal{D}((\Psi^{\bullet(\mathfrak{L}-L)}) \bullet \Phi) = \begin{cases} (l_0, l_1, \dots, l_{L-1}, d) & : \mathfrak{L} = L \\ (l_0, l_1, \dots, l_{L-1}, i, i, \dots, i, d) & : \mathfrak{L} > L. \end{cases} \quad (2.71)$$

This and (2.47) establish (2.66). The proof of Lemma 2.2.11 is thus complete. \square

2.2.4 Parallelizations of ANNs with different lengths

Definition 2.2.12 (Parallelization of ANNs with different length). *Let $n \in \mathbb{N}$, $\Psi = (\Psi_1, \dots, \Psi_n) \in \mathbb{N}^n$ satisfy for all $j \in \{1, 2, \dots, n\}$ that*

$$\mathcal{H}(\Psi_j) = 1 \quad \text{and} \quad \mathcal{I}(\Psi_j) = \mathcal{O}(\Psi_j) \quad (2.72)$$

(cf. Definition 1.3.1). Then we denote by

$$P_{n,\Psi}: \{ \Phi = (\Phi_1, \dots, \Phi_n) \in \mathbb{N}^n : (\forall j \in \{1, 2, \dots, n\} : \mathcal{O}(\Phi_j) = \mathcal{I}(\Psi_j)) \} \rightarrow \mathbf{N} \quad (2.73)$$

the function which satisfies for all $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbb{N}^n$ with $\forall j \in \{1, 2, \dots, n\} : \mathcal{O}(\Phi_j) = \mathcal{I}(\Psi_j)$ that

$$P_{n,\Psi}(\Phi) = \mathbf{P}_n(\mathcal{E}_{\max_{k \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_k), \Psi_1}(\Phi_1), \dots, \mathcal{E}_{\max_{k \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_k), \Psi_n}(\Phi_n)) \quad (2.74)$$

(cf. Definitions 2.2.1 and 2.2.8 and Lemma 2.2.9).

Lemma 2.2.13 (Realizations for parallelizations of ANNs with different length). *Let $a \in C(\mathbb{R}, \mathbb{R})$, $n \in \mathbb{N}$, $\mathbb{I} = (\mathbb{I}_1, \dots, \mathbb{I}_n)$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbb{N}^n$ satisfy for all $j \in \{1, 2, \dots, n\}$, $x \in \mathbb{R}^{\mathcal{O}(\Phi_j)}$ that $\mathcal{H}(\mathbb{I}_j) = 1$, $\mathcal{I}(\mathbb{I}_j) = \mathcal{O}(\mathbb{I}_j) = \mathcal{O}(\Phi_j)$, and $(\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}_j))(x) = x$ (cf. Definitions 1.3.1 and 1.3.4). Then*

(i) it holds that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_{n,\mathbb{I}}(\Phi)) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.75)$$

and

(ii) it holds for all $x_1 \in \mathbb{R}^{\mathcal{I}(\Phi_1)}, x_2 \in \mathbb{R}^{\mathcal{I}(\Phi_2)}, \dots, x_n \in \mathbb{R}^{\mathcal{I}(\Phi_n)}$ that

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_{n,\mathbb{I}}(\Phi)))(x_1, x_2, \dots, x_n) \\ &= ((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n)) \in \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]} \end{aligned} \quad (2.76)$$

(cf. Definition 2.2.12).

Proof of Lemma 2.2.13. Throughout this proof, let $L \in \mathbb{N}$ satisfy $L = \max_{j \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_j)$. Note that item (ii) in Lemma 2.2.9, the assumption that for all $j \in \{1, 2, \dots, n\}$ it holds that $\mathcal{H}(\mathbb{I}_j) = 1$, (2.47), (2.4), and item (ii) in Lemma 2.2.10 demonstrate

(I) that for all $j \in \{1, 2, \dots, n\}$ it holds that $\mathcal{L}(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j)) = L$ and $\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j)) \in C(\mathbb{R}^{\mathcal{I}(\Phi_j)}, \mathbb{R}^{\mathcal{O}(\Phi_j)})$ and

(II) that for all $j \in \{1, 2, \dots, n\}, x \in \mathbb{R}^{\mathcal{I}(\Phi_j)}$ it holds that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j)))(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi_j))(x) \quad (2.77)$$

(cf. Definition 2.2.8). Items (i) and (ii) in Proposition 2.2.3 therefore imply

(A) that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \dots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n))) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.78)$$

and

(B) that for all $x_1 \in \mathbb{R}^{\mathcal{I}(\Phi_1)}, x_2 \in \mathbb{R}^{\mathcal{I}(\Phi_2)}, \dots, x_n \in \mathbb{R}^{\mathcal{I}(\Phi_n)}$ it holds that

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \dots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n))))(x_1, x_2, \dots, x_n) \\ &= \left((\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1)))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_2}(\Phi_2)))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_n}(\Phi_n)))(x_n) \right) \\ &= \left((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n) \right) \end{aligned} \quad (2.79)$$

(cf. Definition 2.2.1). Combining this with (2.74) and the fact that $L = \max_{j \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_j)$ ensures

(C) that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_{n,\mathbb{I}}(\Phi)) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.80)$$

and

(D) that for all $x_1 \in \mathbb{R}^{\mathcal{I}(\Phi_1)}, x_2 \in \mathbb{R}^{\mathcal{I}(\Phi_2)}, \dots, x_n \in \mathbb{R}^{\mathcal{I}(\Phi_n)}$ it holds that

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_{n,\mathbb{I}}(\Phi)))(x_1, x_2, \dots, x_n) \\ &= (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \dots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n))))(x_1, x_2, \dots, x_n) \\ &= \left((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n) \right). \end{aligned} \quad (2.81)$$

This establishes items (i) and (ii). The proof of Lemma 2.2.13 is thus complete. \square

Exercise 2.2.3. For every $d \in \mathbb{N}$ let $F_d: \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfy for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that

$$F_d(x) = (\max\{|x_1|\}, \max\{|x_1|, |x_2|\}, \dots, \max\{|x_1|, |x_2|, \dots, |x_d|\}). \quad (2.82)$$

Prove or disprove the following statement: For all $d \in \mathbb{N}$ there exists $\Phi \in \mathbf{N}$ such that

$$\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\Phi) = F_d \quad (2.83)$$

(cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

2.3 Scalar multiplications of ANNs

2.3.1 Affine transformations as ANNs

Definition 2.3.1 (Affine transformation ANNs). *Let $m, n \in \mathbb{N}$, $W \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^m$. Then we denote by*

$$\mathbf{A}_{W,B} \in (\mathbb{R}^{m \times n} \times \mathbb{R}^m) \subseteq \mathbf{N} \quad (2.84)$$

the ANN given by

$$\mathbf{A}_{W,B} = (W, B) \quad (2.85)$$

(cf. Definitions 1.3.1 and 1.3.2).

Lemma 2.3.2 (Realizations of affine transformation of ANNs). *Let $m, n \in \mathbb{N}$, $W \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^m$. Then*

(i) it holds that $\mathcal{D}(\mathbf{A}_{W,B}) = (n, m) \in \mathbb{N}^2$,

(ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$, and

(iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}))(x) = Wx + B \quad (2.86)$$

(cf. Definitions 1.3.1, 1.3.4, and 2.3.1).

Proof of Lemma 2.3.2. Note that the fact that $\mathbf{A}_{W,B} \in (\mathbb{R}^{m \times n} \times \mathbb{R}^m) \subseteq \mathbf{N}$ implies that

$$\mathcal{D}(\mathbf{A}_{W,B}) = (n, m) \in \mathbb{N}^2. \quad (2.87)$$

This proves item (i). Next observe that the fact that

$$\mathbf{A}_{W,B} = (W, B) \in (\mathbb{R}^{m \times n} \times \mathbb{R}^m) \quad (2.88)$$

and (1.68) demonstrate that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}))(x) = Wx + B. \quad (2.89)$$

This establishes items (ii) and (iii). The proof of Lemma 2.3.2 is thus complete. The proof of Lemma 2.3.2 is thus complete. \square

Lemma 2.3.3 (Compositions with affine transformation ANNs). *Let $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then*

(i) it holds for all $m \in \mathbb{N}$, $W \in \mathbb{R}^{m \times \mathcal{O}(\Phi)}$, $B \in \mathbb{R}^m$ that

$$\mathcal{D}(\mathbf{A}_{W,B} \bullet \Phi) = (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \dots, \mathbb{D}_{\mathcal{H}(\Phi)}(\Phi), m), \quad (2.90)$$

(ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $m \in \mathbb{N}$, $W \in \mathbb{R}^{m \times \mathcal{O}(\Phi)}$, $B \in \mathbb{R}^m$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^m)$,

(iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $m \in \mathbb{N}$, $W \in \mathbb{R}^{m \times \mathcal{O}(\Phi)}$, $B \in \mathbb{R}^m$, $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B} \bullet \Phi))(x) = W((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)) + B, \quad (2.91)$$

(iv) it holds for all $n \in \mathbb{N}$, $W \in \mathbb{R}^{\mathcal{I}(\Phi) \times n}$, $B \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that

$$\mathcal{D}(\Phi \bullet \mathbf{A}_{W,B}) = (n, \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \quad (2.92)$$

(v) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $n \in \mathbb{N}$, $W \in \mathbb{R}^{\mathcal{I}(\Phi) \times n}$, $B \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that $\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^{\mathcal{O}(\Phi)})$, and

(vi) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $n \in \mathbb{N}$, $W \in \mathbb{R}^{\mathcal{I}(\Phi) \times n}$, $B \in \mathbb{R}^{\mathcal{I}(\Phi)}$, $x \in \mathbb{R}^n$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbf{A}_{W,B}))(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))(Wx + B) \quad (2.93)$$

(cf. Definitions 1.3.4, 2.1.1, and 2.3.1).

Proof of Lemma 2.3.3. Note that Lemma 2.3.2 shows that for all $m, n \in \mathbb{N}$, $W \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^m$, $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}))(x) = Wx + B \quad (2.94)$$

(cf. Definitions 1.3.4 and 2.3.1). Combining this and Proposition 2.1.2 proves items (i), (ii), (iii), (iv), (v), and (vi). The proof of Lemma 2.3.3 is thus complete. \square

2.3.2 Scalar multiplications of ANNs

Definition 2.3.4 (Scalar multiplications of ANNs). We denote by $(\cdot) \otimes (\cdot): \mathbb{R} \times \mathbf{N} \rightarrow \mathbf{N}$ the function which satisfies for all $\lambda \in \mathbb{R}$, $\Phi \in \mathbf{N}$ that

$$\lambda \otimes \Phi = \mathbf{A}_{\lambda \mathbf{I}_{\mathcal{O}(\Phi)}, 0} \bullet \Phi \quad (2.95)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, and 3.2.1).

Lemma 2.3.5. Let $\lambda \in \mathbb{R}$, $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then

(i) it holds that $\mathcal{D}(\lambda \otimes \Phi) = \mathcal{D}(\Phi)$,

(ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\lambda \otimes \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$, and

(iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\lambda \otimes \Phi))(x) = \lambda((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)) \quad (2.96)$$

(cf. Definitions 1.3.4 and 2.3.4).

Proof of Lemma 2.3.5. Throughout this proof, let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$ satisfy

$$L = \mathcal{L}(\Phi) \quad \text{and} \quad (l_0, l_1, \dots, l_L) = \mathcal{D}(\Phi). \quad (2.97)$$

Observe that item (i) in Lemma 2.3.2 ensures that

$$\mathcal{D}(\mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)},0}) = (\mathcal{O}(\Phi), \mathcal{O}(\Phi)) \quad (2.98)$$

(cf. Definitions 2.3.1 and 3.2.1). Combining this and item (i) in Lemma 2.3.3 implies that

$$\mathcal{D}(\lambda \otimes \Phi) = \mathcal{D}(\mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)},0} \bullet \Phi) = (l_0, l_1, \dots, l_{L-1}, \mathcal{O}(\Phi)) = \mathcal{D}(\Phi) \quad (2.99)$$

(cf. Definitions 2.1.1 and 2.3.4). This establishes item (i). Note that items (ii) and (iii) in Lemma 2.3.3 demonstrate that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\lambda \otimes \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$ and

$$\begin{aligned} (\mathcal{R}_a^{\mathbf{N}}(\lambda \otimes \Phi))(x) &= (\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)},0} \bullet \Phi))(x) \\ &= \lambda I_{\mathcal{O}(\Phi)}((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)) \\ &= \lambda((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)) \end{aligned} \quad (2.100)$$

(cf. Definition 1.3.4). This proves items (ii) and (iii). The proof of Lemma 2.3.5 is thus complete. \square

2.4 Sums of ANNs with the same length

2.4.1 Sums of vectors as ANNs

Definition 2.4.1 (Sums of vectors as ANNs). *Let $m, n \in \mathbb{N}$. Then we denote by*

$$\mathbb{S}_{m,n} \in (\mathbb{R}^{m \times (mn)} \times \mathbb{R}^m) \subseteq \mathbf{N} \quad (2.101)$$

the ANN given by

$$\mathbb{S}_{m,n} = \mathbf{A}_{(I_m \ I_m \ \dots \ I_m),0} \quad (2.102)$$

(cf. Definitions 1.3.1, 1.3.2, 2.3.1, and 3.2.1).

Lemma 2.4.2. *Let $m, n \in \mathbb{N}$. Then*

(i) it holds that $\mathcal{D}(\mathbb{S}_{m,n}) = (mn, m) \in \mathbb{N}^2$,

(ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$, and

(iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \quad (2.103)$$

(cf. Definitions 1.3.1, 1.3.4, and 2.4.1).

Proof of Lemma 2.4.2. Observe that the fact that $\mathbb{S}_{m,n} \in (\mathbb{R}^{m \times (mn)} \times \mathbb{R}^m)$ shows that

$$\mathcal{D}(\mathbb{S}_{m,n}) = (mn, m) \in \mathbb{N}^2 \quad (2.104)$$

(cf. Definitions 1.3.1 and 2.4.1). This establishes item (i). Note that items (ii) and (iii) in Lemma 2.3.2 ensure that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbb{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$ and

$$\begin{aligned} (\mathcal{R}_a^{\mathbb{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) &= (\mathcal{R}_a^{\mathbb{N}}(\mathbf{A}_{(\mathbb{I}_m \ \mathbb{I}_m \ \dots \ \mathbb{I}_m), 0}))(x_1, x_2, \dots, x_n) \\ &= (\mathbb{I}_m \ \mathbb{I}_m \ \dots \ \mathbb{I}_m)(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \end{aligned} \quad (2.105)$$

(cf. Definitions 1.3.4, 2.3.1, and 3.2.1). This proves items (ii) and (iii). The proof of Lemma 2.4.2 is thus complete. \square

Lemma 2.4.3. *Let $m, n \in \mathbb{N}$, $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbb{N}$ satisfy $\mathcal{O}(\Phi) = mn$ (cf. Definition 1.3.1). Then*

(i) *it holds that $\mathcal{R}_a^{\mathbb{N}}(\mathbb{S}_{m,n} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^m)$ and*

(ii) *it holds for all $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$, $y_1, y_2, \dots, y_n \in \mathbb{R}^m$ with $(\mathcal{R}_a^{\mathbb{N}}(\Phi))(x) = (y_1, y_2, \dots, y_n)$ that*

$$(\mathcal{R}_a^{\mathbb{N}}(\mathbb{S}_{m,n} \bullet \Phi))(x) = \sum_{k=1}^n y_k \quad (2.106)$$

(cf. Definitions 1.3.4, 2.1.1, and 2.4.1).

Proof of Lemma 2.4.3. Observe that Lemma 2.4.2 implies that for all $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbb{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbb{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \quad (2.107)$$

(cf. Definitions 1.3.4 and 2.4.1). Combining this and item (v) in Proposition 2.1.2 establishes items (i) and (ii). The proof of Lemma 2.4.3 is thus complete. \square

Lemma 2.4.4. *Let $n \in \mathbb{N}$, $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbb{N}$ (cf. Definition 1.3.1). Then*

(i) *it holds that $\mathcal{R}_a^{\mathbb{N}}(\Phi \bullet \mathbb{S}_{\mathcal{I}(\Phi), n}) \in C(\mathbb{R}^{n\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$ and*

(ii) it holds for all $x_1, x_2, \dots, x_n \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbb{S}_{\mathcal{I}(\Phi), n})) (x_1, x_2, \dots, x_n) = (\mathcal{R}_a^{\mathbf{N}}(\Phi)) \left(\sum_{k=1}^n x_k \right) \quad (2.108)$$

(cf. Definitions 1.3.4, 2.1.1, and 2.4.1).

Proof of Lemma 2.4.4. Note that Lemma 2.4.2 demonstrates that for all $m \in \mathbb{N}$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m, n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m, n})) (x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \quad (2.109)$$

(cf. Definitions 1.3.4 and 2.4.1). Combining this and item (v) in Proposition 2.1.2 proves items (i) and (ii). The proof of Lemma 2.4.4 is thus complete. \square

2.4.2 Concatenation of vectors as ANNs

Definition 2.4.5 (Transpose of a matrix). Let $m, n \in \mathbb{N}$, $A \in \mathbb{R}^{m \times n}$. Then we denote by $A^* \in \mathbb{R}^{n \times m}$ the transpose of A .

Definition 2.4.6 (Concatenation of vectors as ANNs). Let $m, n \in \mathbb{N}$. Then we denote by

$$\mathbb{T}_{m, n} \in (\mathbb{R}^{(mn) \times m} \times \mathbb{R}^{mn}) \subseteq \mathbf{N} \quad (2.110)$$

the ANN given by

$$\mathbb{T}_{m, n} = \mathbf{A}_{(\mathbb{I}_m \ \mathbb{I}_m \ \dots \ \mathbb{I}_m)^*, 0} \quad (2.111)$$

(cf. Definitions 1.3.1, 1.3.2, 2.3.1, 2.4.5, and 3.2.1).

Lemma 2.4.7. Let $m, n \in \mathbb{N}$. Then

(i) it holds that $\mathcal{D}(\mathbb{T}_{m, n}) = (m, mn) \in \mathbb{N}^2$,

(ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m, n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$, and

(iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^m$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m, n})) (x) = (x, x, \dots, x) \quad (2.112)$$

(cf. Definitions 1.3.1, 1.3.4, and 2.4.6).

Proof of Lemma 2.4.7. Observe that the fact that $\mathbb{T}_{m,n} \in (\mathbb{R}^{(mn) \times m} \times \mathbb{R}^{mn})$ shows that

$$\mathcal{D}(\mathbb{T}_{m,n}) = (m, mn) \in \mathbb{N}^2 \quad (2.113)$$

(cf. Definitions 1.3.1 and 2.4.6). This establishes item (i). Note that item (iii) in Lemma 2.3.2 ensures that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbb{N}}(\mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$ and

$$\begin{aligned} (\mathcal{R}_a^{\mathbb{N}}(\mathbb{T}_{m,n}))(x) &= (\mathcal{R}_a^{\mathbb{N}}(\mathbf{A}_{(\mathbb{I}_m \ \mathbb{I}_m \ \dots \ \mathbb{I}_m)^*, 0}))(x) \\ &= (\mathbb{I}_m \ \mathbb{I}_m \ \dots \ \mathbb{I}_m)^* x = (x, x, \dots, x) \end{aligned} \quad (2.114)$$

(cf. Definitions 1.3.4, 2.3.1, 2.4.5, and 3.2.1). This proves items (ii) and (iii). The proof of Lemma 2.4.7 is thus complete. \square

Lemma 2.4.8. *Let $n \in \mathbb{N}$, $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbb{N}$ (cf. Definition 1.3.1). Then*

(i) *it holds that $\mathcal{R}_a^{\mathbb{N}}(\mathbb{T}_{\mathcal{O}(\Phi), n} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{n\mathcal{O}(\Phi)})$ and*

(ii) *it holds for all $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that*

$$(\mathcal{R}_a^{\mathbb{N}}(\mathbb{T}_{\mathcal{O}(\Phi), n} \bullet \Phi))(x) = ((\mathcal{R}_a^{\mathbb{N}}(\Phi))(x), (\mathcal{R}_a^{\mathbb{N}}(\Phi))(x), \dots, (\mathcal{R}_a^{\mathbb{N}}(\Phi))(x)) \quad (2.115)$$

(cf. Definitions 1.3.4, 2.1.1, and 2.4.6).

Proof of Lemma 2.4.8. Observe that Lemma 2.4.7 implies that for all $m \in \mathbb{N}$, $x \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbb{N}}(\mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$ and

$$(\mathcal{R}_a^{\mathbb{N}}(\mathbb{T}_{m,n}))(x) = (x, x, \dots, x) \quad (2.116)$$

(cf. Definitions 1.3.4 and 2.4.6). Combining this and item (v) in Proposition 2.1.2 establishes items (i) and (ii). The proof of Lemma 2.4.8 is thus complete. \square

Lemma 2.4.9. *Let $m, n \in \mathbb{N}$, $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbb{N}$ satisfy $\mathcal{I}(\Phi) = mn$ (cf. Definition 1.3.1). Then*

(i) *it holds that $\mathcal{R}_a^{\mathbb{N}}(\Phi \bullet \mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{\mathcal{O}(\Phi)})$ and*

(ii) *it holds for all $x \in \mathbb{R}^m$ that*

$$(\mathcal{R}_a^{\mathbb{N}}(\Phi \bullet \mathbb{T}_{m,n}))(x) = (\mathcal{R}_a^{\mathbb{N}}(\Phi))(x, x, \dots, x) \quad (2.117)$$

(cf. Definitions 1.3.4, 2.1.1, and 2.4.6).

Proof of Lemma 2.4.9. Note that Lemma 2.4.7 demonstrates that for all $x \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}))(x) = (x, x, \dots, x) \quad (2.118)$$

(cf. Definitions 1.3.4 and 2.4.6). Combining this and item (v) in Proposition 2.1.2 proves items (i) and (ii). The proof of Lemma 2.4.9 is thus complete. \square

2.4.3 Sums of ANNs

Definition 2.4.10 (Sums of ANNs with the same length). *Let $m \in \mathbb{Z}$, $n \in \{m, m+1, \dots\}$, $\Phi_m, \Phi_{m+1}, \dots, \Phi_n \in \mathbf{N}$ satisfy for all $k \in \{m, m+1, \dots, n\}$ that*

$$\mathcal{L}(\Phi_k) = \mathcal{L}(\Phi_m), \quad \mathcal{I}(\Phi_k) = \mathcal{I}(\Phi_m), \quad \text{and} \quad \mathcal{O}(\Phi_k) = \mathcal{O}(\Phi_m) \quad (2.119)$$

(cf. Definition 1.3.1). Then we denote by $\bigoplus_{k=m}^n \Phi_k \in \mathbf{N}$ (we denote by $\Phi_m \oplus \Phi_{m+1} \oplus \dots \oplus \Phi_n \in \mathbf{N}$) the ANN given by

$$\bigoplus_{k=m}^n \Phi_k = (\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) \in \mathbf{N} \quad (2.120)$$

(cf. Definitions 1.3.2, 2.1.1, 2.2.1, 2.4.1, and 2.4.6).

Lemma 2.4.11 (Realizations of sums of ANNs). *Let $m \in \mathbb{Z}$, $n \in \{m, m+1, \dots\}$, $\Phi_m, \Phi_{m+1}, \dots, \Phi_n \in \mathbf{N}$ satisfy for all $k \in \{m, m+1, \dots, n\}$ that*

$$\mathcal{L}(\Phi_k) = \mathcal{L}(\Phi_m), \quad \mathcal{I}(\Phi_k) = \mathcal{I}(\Phi_m), \quad \text{and} \quad \mathcal{O}(\Phi_k) = \mathcal{O}(\Phi_m) \quad (2.121)$$

(cf. Definition 1.3.1). Then

(i) it holds that $\mathcal{L}(\bigoplus_{k=m}^n \Phi_k) = \mathcal{L}(\Phi_m)$,

(ii) it holds that

$$\mathcal{D}\left(\bigoplus_{k=m}^n \Phi_k\right) = \left(\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{H}(\Phi_m)}(\Phi_k), \mathcal{O}(\Phi_m)\right), \quad (2.122)$$

and

(iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that

$$\mathcal{R}_a^{\mathbf{N}}\left(\bigoplus_{k=m}^n \Phi_k\right) = \sum_{k=m}^n (\mathcal{R}_a^{\mathbf{N}}(\Phi_k)) \quad (2.123)$$

(cf. Definitions 1.3.4 and 2.4.10).

Proof of Lemma 2.4.11. First, observe that Lemma 2.2.2 shows that

$$\begin{aligned}
 & \mathcal{D}(\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)) \\
 &= \left(\sum_{k=m}^n \mathbb{D}_0(\Phi_k), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)}(\Phi_k) \right) \\
 &= \left((n-m+1)\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \right. \\
 & \qquad \qquad \qquad \left. (n-m+1)\mathcal{O}(\Phi_m) \right)
 \end{aligned} \tag{2.124}$$

(cf. Definition 2.2.1). Next note that item (i) in Lemma 2.4.2 ensures that

$$\mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1}) = ((n-m+1)\mathcal{O}(\Phi_m), \mathcal{O}(\Phi_m)) \tag{2.125}$$

(cf. Definition 2.4.1). This, (2.124), and item (i) in Proposition 2.1.2 imply that

$$\begin{aligned}
 & \mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)]) \\
 &= \left((n-m+1)\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m) \right).
 \end{aligned} \tag{2.126}$$

Furthermore, observe that item (i) in Lemma 2.4.7 establishes that

$$\mathcal{D}(\mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) = (\mathcal{I}(\Phi_m), (n-m+1)\mathcal{I}(\Phi_m)) \tag{2.127}$$

(cf. Definitions 2.1.1 and 2.4.6). Combining this, (2.126), and item (i) in Proposition 2.1.2 demonstrates that

$$\begin{aligned}
 & \mathcal{D}\left(\bigoplus_{k=m}^n \Phi_k\right) \\
 &= \mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m), (n-m+1)} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), (n-m+1)}) \\
 &= \left(\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m) \right)
 \end{aligned} \tag{2.128}$$

(cf. Definition 2.4.10). This proves items (i) and (ii). Note that Lemma 2.4.9 and (2.124) show that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi_m)}$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) \in C(\mathbb{R}^{\mathcal{I}(\Phi_m)}, \mathbb{R}^{(n-m+1)\mathcal{O}(\Phi_m)}) \tag{2.129}$$

and

$$\begin{aligned}
 & (\mathcal{R}_a^{\mathbf{N}}([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}))(x) \\
 &= (\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)))(x, x, \dots, x)
 \end{aligned} \tag{2.130}$$

(cf. Definition 1.3.4). Combining this with item (ii) in Proposition 2.2.3 ensures that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi_m)}$ it holds that

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}))(x) \\ &= ((\mathcal{R}_a^{\mathbf{N}}(\Phi_m))(x), (\mathcal{R}_a^{\mathbf{N}}(\Phi_{m+1}))(x), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x)) \in \mathbb{R}^{(n-m+1)\mathcal{O}(\Phi_m)}. \end{aligned} \quad (2.131)$$

Lemma 2.4.3, (2.125), and Lemma 2.1.3 hence imply that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi_m)}$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\bigoplus_{k=m}^n \Phi_k) \in C(\mathbb{R}^{\mathcal{I}(\Phi_m)}, \mathbb{R}^{\mathcal{O}(\Phi_m)})$ and

$$\begin{aligned} & \left(\mathcal{R}_a^{\mathbf{N}} \left(\bigoplus_{k=m}^n \Phi_k \right) \right) (x) \\ &= \left(\mathcal{R}_a^{\mathbf{N}} (\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) \right) (x) \\ &= \sum_{k=m}^n (\mathcal{R}_a^{\mathbf{N}}(\Phi_k))(x). \end{aligned} \quad (2.132)$$

This establishes item (iii). The proof of Lemma 2.4.11 is thus complete. \square

Part II

Approximation

Chapter 3

One-dimensional ANN approximation results

In learning problems ANNs are heavily used with the aim to approximate certain target functions. In this chapter we review basic ReLU ANN approximation results for a class of one-dimensional target functions (see Section 3.3 below). These results rely strongly on linear interpolation properties of ANNs.

3.1 Linear interpolation of one-dimensional functions

3.1.1 On the modulus of continuity

The next notion, modulus of continuity, will be employed in later results to control the approximation error of a function with ANNs. We will also recall some elementary properties of moduli of continuity.

Definition 3.1.1 (Modulus of continuity). *Let $A \subseteq \mathbb{R}$ be a set and let $f: A \rightarrow \mathbb{R}$ be a function. Then we denote by $w_f: [0, \infty] \rightarrow [0, \infty]$ the function which satisfies for all $h \in [0, \infty]$ that*

$$\begin{aligned} w_f(h) &= \sup(\{|f(x) - f(y)|: (x, y) \in A \text{ with } |x - y| \leq h\} \cup \{0\}) \\ &= \sup(\{r \in \mathbb{R}: (\exists x \in A, y \in A \cap [x - h, x + h]): r = |f(x) - f(y)|\} \cup \{0\}) \end{aligned} \tag{3.1}$$

and we call w_f the modulus of continuity of f .

Lemma 3.1.2 (Elementary properties of moduli of continuity). *Let $A \subseteq \mathbb{R}$ be a set and let $f: A \rightarrow \mathbb{R}$ be a function. Then*

(i) *it holds that w_f is non-decreasing,*

(ii) *it holds that f is uniformly continuous if and only if $\lim_{h \searrow 0} w_f(h) = 0$,*

(iii) *it holds that f is globally bounded if and only if $w_f(\infty) < \infty$, and*

(iv) *it holds for all $x, y \in A$ that $|f(x) - f(y)| \leq w_f(|x - y|)$*

(cf. Definition 3.1.1).

Proof of Lemma 3.1.2. Observe that (3.1) proves items (i), (ii), (iii), and (iv). The proof of Lemma 3.1.2 is thus complete. \square

Lemma 3.1.3 (Subadditivity of moduli of continuity). *Let $a \in [-\infty, \infty]$, $b \in [a, \infty]$, let $f: ([a, b] \cap \mathbb{R}) \rightarrow \mathbb{R}$ be a function, and let $h, \mathfrak{h} \in [0, \infty]$. Then*

$$w_f(h + \mathfrak{h}) \leq w_f(h) + w_f(\mathfrak{h}) \quad (3.2)$$

(cf. Definition 3.1.1).

Proof of Lemma 3.1.3. Throughout this proof, assume without loss of generality that $\mathfrak{h} \leq h < \infty$. Note that the fact that for all $x, y \in [a, b] \cap \mathbb{R}$ with $|x - y| \leq h + \mathfrak{h}$ it holds that $[x - h, x + h] \cap [y - \mathfrak{h}, y + \mathfrak{h}] \cap [a, b] \neq \emptyset$ demonstrates that for all $x, y \in [a, b] \cap \mathbb{R}$ with $|x - y| \leq h + \mathfrak{h}$ there exists $z \in [a, b] \cap \mathbb{R}$ such that

$$|x - z| \leq h \quad \text{and} \quad |y - z| \leq \mathfrak{h}. \quad (3.3)$$

Items (i) and (iv) in Lemma 3.1.2 therefore show that for all $x, y \in [a, b] \cap \mathbb{R}$ with $|x - y| \leq h + \mathfrak{h}$ there exists $z \in [a, b] \cap \mathbb{R}$ such that

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f(z)| + |f(y) - f(z)| \\ &\leq w_f(|x - z|) + w_f(|y - z|) \leq w_f(h) + w_f(\mathfrak{h}) \end{aligned} \quad (3.4)$$

(cf. Definition 3.1.1). Combining this with (3.1) ensures that

$$w_f(h + \mathfrak{h}) \leq w_f(h) + w_f(\mathfrak{h}). \quad (3.5)$$

The proof of Lemma 3.1.3 is thus complete. \square

Lemma 3.1.4 (Properties of moduli of continuity of Lipschitz continuous functions). *Let $A \subseteq \mathbb{R}$, $L \in [0, \infty)$, let $f: A \rightarrow \mathbb{R}$ satisfy for all $x, y \in A$ that*

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.6)$$

and let $h \in [0, \infty)$. Then

$$w_f(h) \leq Lh \quad (3.7)$$

(cf. Definition 3.1.1).

Proof of Lemma 3.1.4. Observe that (3.1) and (3.6) imply that

$$\begin{aligned} w_f(h) &= \sup(\{|f(x) - f(y)| \in [0, \infty): (x, y \in A \text{ with } |x - y| \leq h)\} \cup \{0\}) \\ &\leq \sup(\{L|x - y| \in [0, \infty): (x, y \in A \text{ with } |x - y| \leq h)\} \cup \{0\}) \\ &\leq \sup(\{Lh, 0\}) = Lh \end{aligned} \quad (3.8)$$

(cf. Definition 3.1.1). The proof of Lemma 3.1.4 is thus complete. \square

3.1.2 Linear interpolation of one-dimensional functions

In this section we introduce one-dimensional linear interpolation operator and some of its properties.

Definition 3.1.5 (Linear interpolation operator). *Let $K \in \mathbb{N}$, $\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K, f_0, f_1, \dots, f_K \in \mathbb{R}$ satisfy $\mathfrak{x}_0 < \mathfrak{x}_1 < \dots < \mathfrak{x}_K$. Then we denote by*

$$\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f_0, f_1, \dots, f_K}: \mathbb{R} \rightarrow \mathbb{R} \quad (3.9)$$

the function which satisfies for all $k \in \{1, 2, \dots, K\}$, $x \in (-\infty, \mathfrak{x}_0)$, $y \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k)$, $z \in [\mathfrak{x}_K, \infty)$ that

$$(\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f_0, f_1, \dots, f_K})(x) = f_0, \quad (\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f_0, f_1, \dots, f_K})(z) = f_K, \quad (3.10)$$

$$\text{and} \quad (\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f_0, f_1, \dots, f_K})(y) = f_{k-1} + \left(\frac{y - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}\right)(f_k - f_{k-1}). \quad (3.11)$$

Lemma 3.1.6 (Elementary properties of the linear interpolation operator). *Let $K \in \mathbb{N}$, $\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K, f_0, f_1, \dots, f_K \in \mathbb{R}$ satisfy $\mathfrak{x}_0 < \mathfrak{x}_1 < \dots < \mathfrak{x}_K$. Then*

(i) *it holds for all $k \in \{0, 1, \dots, K\}$ that*

$$(\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f_0, f_1, \dots, f_K})(\mathfrak{x}_k) = f_k, \quad (3.12)$$

(ii) it holds for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$ that

$$(\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f_0, f_1, \dots, f_K})(x) = f_{k-1} + \left(\frac{x - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) (f_k - f_{k-1}), \quad (3.13)$$

and

(iii) it holds for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$ that

$$(\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f_0, f_1, \dots, f_K})(x) = \left(\frac{\mathbf{x}_k - x}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) f_{k-1} + \left(\frac{x - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) f_k. \quad (3.14)$$

(cf. Definition 3.1.5).

Proof of Lemma 3.1.6. Note that (3.11) establishes items (i) and (ii). Observe that item (ii) proves that for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$ it holds that

$$\begin{aligned} (\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f_0, f_1, \dots, f_K})(x) &= \left[\left(\frac{\mathbf{x}_k - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) - \left(\frac{x - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) \right] f_{k-1} + \left(\frac{x - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) f_k \\ &= \left(\frac{\mathbf{x}_k - x}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) f_{k-1} + \left(\frac{x - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) f_k. \end{aligned} \quad (3.15)$$

This establishes item (iii). The proof of Lemma 3.1.6 is thus complete. \square

Proposition 3.1.7 (Approximation and continuity properties for the linear interpolation operator). *Let $K \in \mathbb{N}$, $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K \in \mathbb{R}$ satisfy $\mathbf{x}_0 < \mathbf{x}_1 < \dots < \mathbf{x}_K$ and let $f: [\mathbf{x}_0, \mathbf{x}_K] \rightarrow \mathbb{R}$ be a function. Then*

(i) it holds for all $x, y \in \mathbb{R}$ with $x \neq y$ that

$$\begin{aligned} & \left| (\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f(\mathbf{x}_0), f(\mathbf{x}_1), \dots, f(\mathbf{x}_K)})(x) - (\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f(\mathbf{x}_0), f(\mathbf{x}_1), \dots, f(\mathbf{x}_K)})(y) \right| \\ & \leq \left(\max_{k \in \{1, 2, \dots, K\}} \left(\frac{w_f(\mathbf{x}_k - \mathbf{x}_{k-1})}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) \right) |x - y| \end{aligned} \quad (3.16)$$

and

(ii) it holds that

$$\sup_{x \in [\mathbf{x}_0, \mathbf{x}_K]} \left| (\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f(\mathbf{x}_0), f(\mathbf{x}_1), \dots, f(\mathbf{x}_K)})(x) - f(x) \right| \leq w_f(\max_{k \in \{1, 2, \dots, K\}} |\mathbf{x}_k - \mathbf{x}_{k-1}|) \quad (3.17)$$

(cf. Definitions 3.1.1 and 3.1.5).

Proof of Proposition 3.1.7. Throughout this proof, let $L \in [0, \infty]$ satisfy

$$L = \max_{k \in \{1, 2, \dots, K\}} \left(\frac{w_f(\mathbf{x}_k - \mathbf{x}_{k-1})}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) \quad (3.18)$$

and let $\mathfrak{l}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that

$$\mathfrak{l}(x) = (\mathcal{L}_{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_K}^{f(\mathbf{x}_0), f(\mathbf{x}_1), \dots, f(\mathbf{x}_K)})(x) \quad (3.19)$$

(cf. Definitions 3.1.1 and 3.1.5). Observe that item (ii) in Lemma 3.1.6, item (iv) in Lemma 3.1.2, and (3.18) demonstrate that for all $k \in \{1, 2, \dots, K\}$, $x, y \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$ with $x \neq y$ it holds that

$$\begin{aligned} |\mathfrak{l}(x) - \mathfrak{l}(y)| &= \left| \left(\frac{x - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) (f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})) - \left(\frac{y - \mathbf{x}_{k-1}}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) (f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})) \right| \\ &= \left| \left(\frac{f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) (x - y) \right| \leq \left(\frac{w_f(\mathbf{x}_k - \mathbf{x}_{k-1})}{\mathbf{x}_k - \mathbf{x}_{k-1}} \right) |x - y| \leq L|x - y|. \end{aligned} \quad (3.20)$$

Next note that the triangle inequality and item (i) in Lemma 3.1.6 show that for all $k, l \in \{1, 2, \dots, K\}$, $x \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$, $y \in [\mathbf{x}_{l-1}, \mathbf{x}_l]$ with $k < l$ it holds that

$$\begin{aligned} |\mathfrak{l}(x) - \mathfrak{l}(y)| &\leq |\mathfrak{l}(x) - \mathfrak{l}(\mathbf{x}_k)| + |\mathfrak{l}(\mathbf{x}_k) - \mathfrak{l}(\mathbf{x}_{l-1})| + |\mathfrak{l}(\mathbf{x}_{l-1}) - \mathfrak{l}(y)| \\ &= |\mathfrak{l}(x) - \mathfrak{l}(\mathbf{x}_k)| + |f(\mathbf{x}_k) - f(\mathbf{x}_{l-1})| + |\mathfrak{l}(\mathbf{x}_{l-1}) - \mathfrak{l}(y)| \\ &\leq |\mathfrak{l}(x) - \mathfrak{l}(\mathbf{x}_k)| + \left(\sum_{j=k+1}^{l-1} |f(\mathbf{x}_j) - f(\mathbf{x}_{j-1})| \right) + |\mathfrak{l}(\mathbf{x}_{l-1}) - \mathfrak{l}(y)|. \end{aligned} \quad (3.21)$$

Item (iv) in Lemma 3.1.2, and (3.18) hence ensure that for all $k, l \in \{1, 2, \dots, K\}$, $x \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$, $y \in [\mathbf{x}_{l-1}, \mathbf{x}_l]$ with $k < l$ and $x \neq y$ it holds that

$$\begin{aligned} &|\mathfrak{l}(x) - \mathfrak{l}(y)| \\ &\leq |\mathfrak{l}(x) - \mathfrak{l}(\mathbf{x}_k)| + \left(\sum_{j=k+1}^{l-1} w_f(|\mathbf{x}_j - \mathbf{x}_{j-1}|) \right) + |\mathfrak{l}(\mathbf{x}_{l-1}) - \mathfrak{l}(y)| \\ &= |\mathfrak{l}(x) - \mathfrak{l}(\mathbf{x}_k)| + \left(\sum_{j=k+1}^{l-1} \left(\frac{w_f(\mathbf{x}_j - \mathbf{x}_{j-1})}{\mathbf{x}_j - \mathbf{x}_{j-1}} \right) (\mathbf{x}_j - \mathbf{x}_{j-1}) \right) + |\mathfrak{l}(\mathbf{x}_{l-1}) - \mathfrak{l}(y)| \\ &\leq |\mathfrak{l}(\mathbf{x}_k) - \mathfrak{l}(x)| + L \left(\sum_{j=k+1}^{l-1} (\mathbf{x}_j - \mathbf{x}_{j-1}) \right) + |\mathfrak{l}(y) - \mathfrak{l}(\mathbf{x}_{l-1})|. \end{aligned} \quad (3.22)$$

This and (3.21) imply that for all $k, l \in \{1, 2, \dots, K\}$, $x \in [\mathbf{x}_{k-1}, \mathbf{x}_k]$, $y \in [\mathbf{x}_{l-1}, \mathbf{x}_l]$ with $k < l$ and $x \neq y$ it holds that

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| \leq L((\mathbf{x}_k - x) + (\mathbf{x}_{l-1} - \mathbf{x}_k) + (y - \mathbf{x}_{l-1})) = L|x - y|. \quad (3.23)$$

Combining this and (3.20) proves that for all $x, y \in [\mathfrak{r}_0, \mathfrak{r}_K]$ with $x \neq y$ it holds that

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| \leq L|x - y|. \quad (3.24)$$

This, the fact that for all $x, y \in (-\infty, \mathfrak{r}_0]$ with $x \neq y$ it holds that

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| = 0 \leq L|x - y|, \quad (3.25)$$

the fact that for all $x, y \in [\mathfrak{r}_K, \infty)$ with $x \neq y$ it holds that

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| = 0 \leq L|x - y|, \quad (3.26)$$

and the triangle inequality therefore establish that for all $x, y \in \mathbb{R}$ with $x \neq y$ it holds that

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| \leq L|x - y|. \quad (3.27)$$

This proves item (i). Observe that item (iii) in Lemma 3.1.6 demonstrates that for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathfrak{r}_{k-1}, \mathfrak{r}_k]$ it holds that

$$\begin{aligned} |\mathfrak{l}(x) - f(x)| &= \left| \left(\frac{\mathfrak{r}_k - x}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) f(\mathfrak{r}_{k-1}) + \left(\frac{x - \mathfrak{r}_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) f(\mathfrak{r}_k) - f(x) \right| \\ &= \left| \left(\frac{\mathfrak{r}_k - x}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) (f(\mathfrak{r}_{k-1}) - f(x)) + \left(\frac{x - \mathfrak{r}_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) (f(\mathfrak{r}_k) - f(x)) \right| \\ &\leq \left(\frac{\mathfrak{r}_k - x}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) |f(\mathfrak{r}_{k-1}) - f(x)| + \left(\frac{x - \mathfrak{r}_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) |f(\mathfrak{r}_k) - f(x)|. \end{aligned} \quad (3.28)$$

Combining this with (3.1) and Lemma 3.1.2 shows that for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathfrak{r}_{k-1}, \mathfrak{r}_k]$ it holds that

$$\begin{aligned} |\mathfrak{l}(x) - f(x)| &\leq w_f(|\mathfrak{r}_k - \mathfrak{r}_{k-1}|) \left(\frac{\mathfrak{r}_k - x}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} + \frac{x - \mathfrak{r}_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} \right) \\ &= w_f(|\mathfrak{r}_k - \mathfrak{r}_{k-1}|) \leq w_f(\max_{j \in \{1, 2, \dots, K\}} |\mathfrak{r}_j - \mathfrak{r}_{j-1}|). \end{aligned} \quad (3.29)$$

This establishes item (ii). The proof of Proposition 3.1.7 is thus complete. \square

Corollary 3.1.8 (Approximation and Lipschitz continuity properties for the linear interpolation operator). *Let $K \in \mathbb{N}$, $L, \mathfrak{r}_0, \mathfrak{r}_1, \dots, \mathfrak{r}_K \in \mathbb{R}$ satisfy $\mathfrak{r}_0 < \mathfrak{r}_1 < \dots < \mathfrak{r}_K$ and let $f: [\mathfrak{r}_0, \mathfrak{r}_K] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [\mathfrak{r}_0, \mathfrak{r}_K]$ that*

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.30)$$

Then

(i) it holds for all $x, y \in \mathbb{R}$ that

$$\left| (\mathcal{L}_{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K}^{f(\mathbf{r}_0), f(\mathbf{r}_1), \dots, f(\mathbf{r}_K)})(x) - (\mathcal{L}_{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K}^{f(\mathbf{r}_0), f(\mathbf{r}_1), \dots, f(\mathbf{r}_K)})(y) \right| \leq L|x - y| \quad (3.31)$$

and

(ii) it holds that

$$\sup_{x \in [\mathbf{r}_0, \mathbf{r}_K]} \left| (\mathcal{L}_{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K}^{f(\mathbf{r}_0), f(\mathbf{r}_1), \dots, f(\mathbf{r}_K)})(x) - f(x) \right| \leq L \left(\max_{k \in \{1, 2, \dots, K\}} |\mathbf{r}_k - \mathbf{r}_{k-1}| \right) \quad (3.32)$$

(cf. Definition 3.1.5).

Proof of Corollary 3.1.8. Note that Lemma 3.1.4, the assumption that for all $x, y \in [\mathbf{r}_0, \mathbf{r}_K]$ it holds that $|f(x) - f(y)| \leq L|x - y|$, and item (i) in Proposition 3.1.7 ensure that for all $x, y \in \mathbb{R}$ it holds that

$$\begin{aligned} & \left| (\mathcal{L}_{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K}^{f(\mathbf{r}_0), f(\mathbf{r}_1), \dots, f(\mathbf{r}_K)})(x) - (\mathcal{L}_{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K}^{f(\mathbf{r}_0), f(\mathbf{r}_1), \dots, f(\mathbf{r}_K)})(y) \right| \\ & \leq \left(\max_{k \in \{1, 2, \dots, K\}} \left(\frac{L|\mathbf{r}_k - \mathbf{r}_{k-1}|}{|\mathbf{r}_k - \mathbf{r}_{k-1}|} \right) \right) |x - y| = L|x - y|. \end{aligned} \quad (3.33)$$

This proves item (i). Next observe that the assumption that for all $x, y \in [\mathbf{r}_0, \mathbf{r}_K]$ it holds that $|f(x) - f(y)| \leq L|x - y|$, Lemma 3.1.4, and item (ii) in Proposition 3.1.7 imply that

$$\begin{aligned} \sup_{x \in [\mathbf{r}_0, \mathbf{r}_K]} \left| (\mathcal{L}_{\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K}^{f(\mathbf{r}_0), f(\mathbf{r}_1), \dots, f(\mathbf{r}_K)})(x) - f(x) \right| & \leq w_f \left(\max_{k \in \{1, 2, \dots, K\}} |\mathbf{r}_k - \mathbf{r}_{k-1}| \right) \\ & \leq L \left(\max_{k \in \{1, 2, \dots, K\}} |\mathbf{r}_k - \mathbf{r}_{k-1}| \right). \end{aligned} \quad (3.34)$$

This establishes item (ii). The proof of Corollary 3.1.8 is thus complete. \square

3.2 Linear interpolation with ANNs

3.2.1 Activation functions as ANNs

Definition 3.2.1 (Identity matrices). *Let $d \in \mathbb{N}$. Then we denote by $I_d \in \mathbb{R}^{d \times d}$ the identity matrix in $\mathbb{R}^{d \times d}$.*

Definition 3.2.2 (Activation functions as ANNs). *Let $n \in \mathbb{N}$. Then we denote by*

$$\mathbf{i}_n \in ((\mathbb{R}^{n \times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n \times n} \times \mathbb{R}^n)) \subseteq \mathbf{N} \quad (3.35)$$

the ANN given by

$$\mathbf{i}_n = ((\mathbf{I}_n, 0), (\mathbf{I}_n, 0)) \quad (3.36)$$

(cf. Definitions 1.3.1 and 3.2.1).

Lemma 3.2.3 (Realization functions of activation ANNs). *Let $n \in \mathbb{N}$. Then*

(i) it holds that $\mathcal{D}(\mathbf{i}_n) = (n, n, n) \in \mathbb{N}^3$ and

(ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n) = \mathfrak{M}_{a,n} \quad (3.37)$$

(cf. Definitions 1.2.1, 1.3.1, 1.3.4, and 3.2.2).

Proof of Lemma 3.2.3. Observe that the fact that $\mathbf{i}_n \in ((\mathbb{R}^{n \times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n \times n} \times \mathbb{R}^n)) \subseteq \mathbf{N}$ demonstrates that

$$\mathcal{D}(\mathbf{i}_n) = (n, n, n) \in \mathbb{N}^3 \quad (3.38)$$

(cf. Definitions 1.3.1 and 3.2.2). This proves item (i). Note that (1.68) and the fact that

$$\mathbf{i}_n = ((\mathbf{I}_n, 0), (\mathbf{I}_n, 0)) \in ((\mathbb{R}^{n \times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n \times n} \times \mathbb{R}^n)) \quad (3.39)$$

show that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n) \in C(\mathbb{R}^n, \mathbb{R}^n)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n))(x) = \mathbf{I}_n(\mathfrak{M}_{a,n}(\mathbf{I}_n x + 0)) + 0 = \mathfrak{M}_{a,n}(x). \quad (3.40)$$

This establishes item (ii). The proof of Lemma 3.2.3 is thus complete. \square

Lemma 3.2.4 (Compositions of activation ANNs with general ANNs). *Let $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then*

(i) it holds that

$$\begin{aligned} & \mathcal{D}(\mathbf{i}_{\mathcal{O}(\Phi)} \bullet \Phi) \\ &= (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)-1}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)) \in \mathbb{N}^{\mathcal{L}(\Phi)+2}, \end{aligned} \quad (3.41)$$

- (ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_{\mathcal{O}(\Phi)} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$,
- (iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_{\mathcal{O}(\Phi)} \bullet \Phi) = \mathfrak{M}_{a, \mathcal{O}(\Phi)} \circ (\mathcal{R}_a^{\mathbf{N}}(\Phi))$,
- (iv) it holds that

$$\begin{aligned} & \mathcal{D}(\Phi \bullet \mathbf{i}_{\mathcal{I}(\Phi)}) \\ &= (\mathbb{D}_0(\Phi), \mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)-1}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)) \in \mathbb{N}^{\mathcal{L}(\Phi)+2}, \end{aligned} \quad (3.42)$$

- (v) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbf{i}_{\mathcal{I}(\Phi)}) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$, and
- (vi) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbf{i}_{\mathcal{I}(\Phi)}) = (\mathcal{R}_a^{\mathbf{N}}(\Phi)) \circ \mathfrak{M}_{a, \mathcal{I}(\Phi)}$
- (cf. Definitions 1.2.1, 1.3.4, 2.1.1, and 3.2.2).

Proof of Lemma 3.2.4. Observe that Lemma 3.2.3 ensures that for all $n \in \mathbb{N}$, $a \in C(\mathbb{R}, \mathbb{R})$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n) = \mathfrak{M}_{a, n} \quad (3.43)$$

(cf. Definitions 1.2.1, 1.3.4, and 3.2.2). Combining this and Proposition 2.1.2 proves items (i), (ii), (iii), (iv), (v), and (vi). The proof of Lemma 3.2.4 is thus complete. \square

3.2.2 Representations for ReLU ANNs with one hidden neuron

Lemma 3.2.5. *Let $\alpha, \beta, h \in \mathbb{R}$, $\mathbf{H} \in \mathbf{N}$ satisfy*

$$\mathbf{H} = h \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}) \quad (3.44)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, and 3.2.2). Then

- (i) it holds that $\mathbf{H} = ((\alpha, \beta), (h, 0))$,
- (ii) it holds that $\mathcal{D}(\mathbf{H}) = (1, 1, 1) \in \mathbb{N}^3$,
- (iii) it holds that $\mathcal{R}_{\mathbf{t}}^{\mathbf{N}}(\mathbf{H}) \in C(\mathbb{R}, \mathbb{R})$, and
- (iv) it holds for all $x \in \mathbb{R}$ that $(\mathcal{R}_{\mathbf{t}}^{\mathbf{N}}(\mathbf{H}))(x) = h \max\{\alpha x + \beta, 0\}$

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Lemma 3.2.5. Note that Lemma 2.3.2 implies that

$$\mathbf{A}_{\alpha, \beta} = (\alpha, \beta), \quad \mathcal{D}(\mathbf{A}_{\alpha, \beta}) = (1, 1) \in \mathbb{N}^2, \quad \mathcal{R}_{\mathbf{t}}^{\mathbf{N}}(\mathbf{A}_{\alpha, \beta}) \in C(\mathbb{R}, \mathbb{R}), \quad (3.45)$$

and $\forall x \in \mathbb{R}$: $(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{A}_{\alpha,\beta}))(x) = \alpha x + \beta$ (cf. Definitions 1.2.4 and 1.3.4). Proposition 2.1.2, Lemma 3.2.3, Lemma 3.2.4, (1.27), (1.68), and (2.2) hence demonstrate that

$$\mathbf{i}_1 \bullet \mathbf{A}_{\alpha,\beta} = ((\alpha, \beta), (1, 0)), \quad \mathcal{D}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) = (1, 1, 1) \in \mathbb{N}^3, \quad \mathcal{R}_\tau^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) \in C(\mathbb{R}, \mathbb{R}), \quad (3.46)$$

$$\text{and } \forall x \in \mathbb{R}: (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha,\beta}))(x) = \tau(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{A}_{\alpha,\beta}))(x) = \max\{\alpha x + \beta, 0\}. \quad (3.47)$$

This, Lemma 2.3.5, and (2.95) show that

$$\mathbf{H} = h \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) = ((\alpha, \beta), (h, 0)), \quad \mathcal{D}(\mathbf{H}) = (1, 1, 1), \quad \mathcal{R}_\tau^{\mathbf{N}}(\mathbf{H}) \in C(\mathbb{R}, \mathbb{R}), \quad (3.48)$$

$$\text{and } (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{H}))(x) = h((\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha,\beta}))(x)) = h \max\{\alpha x + \beta, 0\}. \quad (3.49)$$

This establishes items (i), (ii), (iii), and (iv). The proof of Lemma 3.2.5 is thus complete. \square

3.2.3 ReLU ANN representations for linear interpolations

Proposition 3.2.6 (ReLU ANN representations for linear interpolations). *Let $K \in \mathbb{N}$, $f_0, f_1, \dots, f_K, \mathfrak{r}_0, \mathfrak{r}_1, \dots, \mathfrak{r}_K \in \mathbb{R}$ satisfy $\mathfrak{r}_0 < \mathfrak{r}_1 < \dots < \mathfrak{r}_K$ and let $\mathbf{F} \in \mathbf{N}$ satisfy*

$$\mathbf{F} = \mathbf{A}_{1,f_0} \bullet \left(\bigoplus_{k=0}^K \left(\left(\frac{(f_{\min\{k+1,K\}} - f_k)}{(\mathfrak{r}_{\min\{k+1,K\}} - \mathfrak{r}_{\min\{k,K-1\}})} - \frac{(f_k - f_{\max\{k-1,0\}})}{(\mathfrak{r}_{\max\{k,1\}} - \mathfrak{r}_{\max\{k-1,0\}})} \right) \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{r}_k}) \right) \right) \quad (3.50)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.2). Then

(i) it holds that $\mathcal{D}(\mathbf{F}) = (1, K + 1, 1) \in \mathbb{N}^3$,

(ii) it holds that $\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) = \mathcal{L}_{\mathfrak{r}_0, \mathfrak{r}_1, \dots, \mathfrak{r}_K}^{f_0, f_1, \dots, f_K}$, and

(iii) it holds that $\mathcal{P}(\mathbf{F}) = 3K + 4$

(cf. Definitions 1.2.4, 1.3.4, and 3.1.5).

Proof of Proposition 3.2.6. Throughout this proof, let $c_0, c_1, \dots, c_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that

$$c_k = \frac{(f_{\min\{k+1,K\}} - f_k)}{(\mathfrak{r}_{\min\{k+1,K\}} - \mathfrak{r}_{\min\{k,K-1\}})} - \frac{(f_k - f_{\max\{k-1,0\}})}{(\mathfrak{r}_{\max\{k,1\}} - \mathfrak{r}_{\max\{k-1,0\}})} \quad (3.51)$$

and let $\Phi_0, \Phi_1, \dots, \Phi_K \in ((\mathbb{R}^{1 \times 1} \times \mathbb{R}^1) \times (\mathbb{R}^{1 \times 1} \times \mathbb{R}^1)) \subseteq \mathbf{N}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that

$$\Phi_k = c_k \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{r}_k}). \quad (3.52)$$

Observe that Lemma 3.2.5 ensures that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$\mathcal{R}_\tau^{\mathbf{N}}(\Phi_k) \in C(\mathbb{R}, \mathbb{R}), \quad \mathcal{D}(\Phi_k) = (1, 1, 1) \in \mathbb{N}^3, \quad (3.53)$$

$$\text{and} \quad \forall x \in \mathbb{R}: (\mathcal{R}_\tau^{\mathbf{N}}(\Phi_k))(x) = c_k \max\{x - \mathfrak{r}_k, 0\} \quad (3.54)$$

(cf. Definitions 1.2.4 and 1.3.4). This, Lemma 2.3.3, Lemma 2.4.11, and (3.50) prove that

$$\mathcal{D}(\mathbf{F}) = (1, K + 1, 1) \in \mathbb{N}^3 \quad \text{and} \quad \mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}). \quad (3.55)$$

This establishes item (i). Note that item (i) and (1.55) imply that

$$\mathcal{P}(\mathbf{F}) = 2(K + 1) + (K + 2) = 3K + 4. \quad (3.56)$$

This demonstrates item (iii). Observe that (3.51), (3.54), Lemma 2.3.3, and Lemma 2.4.11 show that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) = f_0 + \sum_{k=0}^K (\mathcal{R}_\tau^{\mathbf{N}}(\Phi_k))(x) = f_0 + \sum_{k=0}^K c_k \max\{x - \mathfrak{r}_k, 0\}. \quad (3.57)$$

This and the fact that for all $k \in \{0, 1, \dots, K\}$ it holds that $\mathfrak{r}_0 \leq \mathfrak{r}_k$ ensure that for all $x \in (-\infty, \mathfrak{r}_0]$ it holds that

$$(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) = f_0 + 0 = f_0. \quad (3.58)$$

Next we claim that for all $k \in \{1, 2, \dots, K\}$ it holds that

$$\sum_{n=0}^{k-1} c_n = \frac{f_k - f_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}}. \quad (3.59)$$

We now prove (3.59) by induction on $k \in \{1, 2, \dots, K\}$. For the base case $k = 1$ observe that (3.51) proves that

$$\sum_{n=0}^0 c_n = c_0 = \frac{f_1 - f_0}{\mathfrak{r}_1 - \mathfrak{r}_0}. \quad (3.60)$$

This establishes (3.59) in the base case $k = 1$. For the induction step note that (3.51) implies that for all $k \in \mathbb{N} \cap (1, \infty) \cap (0, K]$ with $\sum_{n=0}^{k-2} c_n = \frac{f_{k-1} - f_{k-2}}{\mathfrak{r}_{k-1} - \mathfrak{r}_{k-2}}$ it holds that

$$\sum_{n=0}^{k-1} c_n = c_{k-1} + \sum_{n=0}^{k-2} c_n = \frac{f_k - f_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}} - \frac{f_{k-1} - f_{k-2}}{\mathfrak{r}_{k-1} - \mathfrak{r}_{k-2}} + \frac{f_{k-1} - f_{k-2}}{\mathfrak{r}_{k-1} - \mathfrak{r}_{k-2}} = \frac{f_k - f_{k-1}}{\mathfrak{r}_k - \mathfrak{r}_{k-1}}. \quad (3.61)$$

Induction thus demonstrates (3.59). Next observe that (3.57), (3.59), and the fact that for all $k \in \{1, 2, \dots, K\}$ it holds that $\mathfrak{r}_{k-1} < \mathfrak{r}_k$ show that for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathfrak{r}_{k-1}, \mathfrak{r}_k]$

it holds that

$$\begin{aligned}
 (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) &= \sum_{n=0}^K c_n (\max\{x - \mathfrak{x}_n, 0\} - \max\{\mathfrak{x}_{k-1} - \mathfrak{x}_n, 0\}) \\
 &= \sum_{n=0}^{k-1} c_n [(x - \mathfrak{x}_n) - (\mathfrak{x}_{k-1} - \mathfrak{x}_n)] = \sum_{n=0}^{k-1} c_n (x - \mathfrak{x}_{k-1}) \quad (3.62) \\
 &= \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}).
 \end{aligned}$$

Next we claim that for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ it holds that

$$(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) = f_{k-1} + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}). \quad (3.63)$$

We now prove (3.63) by induction on $k \in \{1, 2, \dots, K\}$. For the base case $k = 1$ note that (3.58) and (3.62) ensure that for all $x \in [\mathfrak{x}_0, \mathfrak{x}_1]$ it holds that

$$(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) = (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_0) + (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_0) = f_0 + \left(\frac{f_1 - f_0}{\mathfrak{x}_1 - \mathfrak{x}_0} \right) (x - \mathfrak{x}_0). \quad (3.64)$$

This proves (3.63) in the base case $k = 1$. For the induction step observe that (3.62) establishes that for all $k \in \mathbb{N} \cap (1, \infty) \cap [1, K]$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ with $\forall y \in [\mathfrak{x}_{k-2}, \mathfrak{x}_{k-1}]$: $(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(y) = f_{k-2} + \left(\frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} \right) (y - \mathfrak{x}_{k-2})$ it holds that

$$\begin{aligned}
 (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) &= (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) + (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) \\
 &= f_{k-2} + \left(\frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} \right) (\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}) + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}) \quad (3.65) \\
 &= f_{k-1} + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}).
 \end{aligned}$$

Induction thus implies (3.63). Furthermore, note that (3.51) and (3.59) demonstrate that

$$\sum_{n=0}^K c_n = c_K + \sum_{n=0}^{K-1} c_n = -\frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} + \frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} = 0. \quad (3.66)$$

The fact that for all $k \in \{0, 1, \dots, K\}$ it holds that $\mathfrak{x}_k \leq \mathfrak{x}_K$ and (3.57) therefore show that for all $x \in [\mathfrak{x}_K, \infty)$ it holds that

$$\begin{aligned}
 (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_K) &= \left[\sum_{n=0}^K c_n (\max\{x - \mathfrak{x}_n, 0\} - \max\{\mathfrak{x}_K - \mathfrak{x}_n, 0\}) \right] \\
 &= \sum_{n=0}^K c_n [(x - \mathfrak{x}_n) - (\mathfrak{x}_K - \mathfrak{x}_n)] = \sum_{n=0}^K c_n (x - \mathfrak{x}_K) = 0. \quad (3.67)
 \end{aligned}$$

This and (3.63) ensure that for all $x \in [\mathfrak{x}_K, \infty)$ it holds that

$$(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) = (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_K) = f_{K-1} + \left(\frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}}\right)(\mathfrak{x}_K - \mathfrak{x}_{K-1}) = f_K. \quad (3.68)$$

Combining this, (3.58), (3.63), and (3.11) proves item (ii). The proof of Proposition 3.2.6 is thus complete. \square

Exercise 3.2.1. Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that $\mathcal{P}(\Phi) \leq 16$ and

$$\sup_{x \in [-2\pi, 2\pi]} |\cos(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\Phi))(x)| \leq \frac{1}{2} \quad (3.69)$$

(cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Exercise 3.2.2. Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that $\mathcal{I}(\Phi) = 4$, $\mathcal{O}(\Phi) = 1$, $\mathcal{P}(\Phi) \leq 60$, and $\forall x, y, u, v \in \mathbb{R}: (\mathcal{R}_\tau^{\mathbf{N}}(\Phi))(x, y, u, v) = \max\{x, y, u, v\}$ (cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Exercise 3.2.3. Prove or disprove the following statement: For every $m \in \mathbb{N}$ there exists $\Phi \in \mathbf{N}$ such that $\mathcal{I}(\Phi) = 2^m$, $\mathcal{O}(\Phi) = 1$, $\mathcal{P}(\Phi) \leq 3(2^m(2^m + 1))$, and $\forall x = (x_1, x_2, \dots, x_{2^m}) \in \mathbb{R}: (\mathcal{R}_\tau^{\mathbf{N}}(\Phi))(x) = \max\{x_1, x_2, \dots, x_{2^m}\}$ (cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

3.3 ANN approximations results for one-dimensional functions

3.3.1 Constructive ANN approximation results

Proposition 3.3.1 (ANN approximations through linear interpolations). *Let $K \in \mathbb{N}$, $L, a, \mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K \in \mathbb{R}$, $b \in (a, \infty)$ satisfy for all $k \in \{0, 1, \dots, K\}$ that $\mathfrak{x}_k = a + \frac{k(b-a)}{K}$, let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.70)$$

and let $\mathbf{F} \in \mathbf{N}$ satisfy

$$\mathbf{F} = \mathbf{A}_{1, f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^K \left(\left(\frac{K(f(\mathfrak{x}_{\min\{k+1, K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1, 0\}}))}{(b-a)} \right) \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1, -\mathfrak{x}_k}) \right) \right) \quad (3.71)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.2). Then

- (i) it holds that $\mathcal{D}(\mathbf{F}) = (1, K + 1, 1)$,
- (ii) it holds that $\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) = \mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)}$,
- (iii) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,

(iv) it holds that $\sup_{x \in [a, b]} |(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq L(b-a)K^{-1}$, and

(v) it holds that $\mathcal{P}(\mathbf{F}) = 3K + 4$

(cf. Definitions 1.2.4, 1.3.4, and 3.1.5).

Proof of Proposition 3.3.1. Observe that the fact that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$\mathfrak{r}_{\min\{k+1, K\}} - \mathfrak{r}_{\min\{k, K-1\}} = \mathfrak{r}_{\max\{k, 1\}} - \mathfrak{r}_{\max\{k-1, 0\}} = (b-a)K^{-1} \quad (3.72)$$

establishes that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$\begin{aligned} & \frac{(f(\mathfrak{r}_{\min\{k+1, K\}}) - f(\mathfrak{r}_k))}{(\mathfrak{r}_{\min\{k+1, K\}} - \mathfrak{r}_{\min\{k, K-1\}})} - \frac{(f(\mathfrak{r}_k) - f(\mathfrak{r}_{\max\{k-1, 0\}}))}{(\mathfrak{r}_{\max\{k, 1\}} - \mathfrak{r}_{\max\{k-1, 0\}})} \\ &= \frac{K(f(\mathfrak{r}_{\min\{k+1, K\}}) - 2f(\mathfrak{r}_k) + f(\mathfrak{r}_{\max\{k-1, 0\}}))}{(b-a)}. \end{aligned} \quad (3.73)$$

This and Proposition 3.2.6 prove items (i), (ii), and (v). Note that item (i) in Corollary 3.1.8, item (ii), and the assumption that for all $x, y \in [a, b]$ it holds that

$$|f(x) - f(y)| \leq L|x - y| \quad (3.74)$$

prove item (iii). Observe that item (ii), the assumption that for all $x, y \in [a, b]$ it holds that

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.75)$$

item (ii) in Corollary 3.1.8, and the fact that for all $k \in \{1, 2, \dots, K\}$ it holds that

$$\mathfrak{r}_k - \mathfrak{r}_{k-1} = \frac{(b-a)}{K} \quad (3.76)$$

imply that for all $x \in [a, b]$ it holds that

$$|(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq L \left(\max_{k \in \{1, 2, \dots, K\}} |\mathfrak{r}_k - \mathfrak{r}_{k-1}| \right) = \frac{L(b-a)}{K}. \quad (3.77)$$

This establishes item (iv). The proof of Proposition 3.3.1 is thus complete. \square

Lemma 3.3.2 (Approximations through ANNs with constant realizations). *Let $L, a \in \mathbb{R}$, $b \in [a, \infty)$, $\xi \in [a, b]$, let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.78)$$

and let $\mathbf{F} \in \mathbf{N}$ satisfy

$$\mathbf{F} = \mathbf{A}_{1,f(\xi)} \bullet (0 \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\xi})) \quad (3.79)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, and 3.2.2). Then

(i) it holds that $\mathcal{D}(\mathbf{F}) = (1, 1, 1)$,

(ii) it holds that $\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,

(iii) it holds for all $x \in \mathbb{R}$ that $(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) = f(\xi)$,

(iv) it holds that $\sup_{x \in [a,b]} |(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq L \max\{\xi - a, b - \xi\}$, and

(v) it holds that $\mathcal{P}(\mathbf{F}) = 4$

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Lemma 3.3.2. Note that items (i) and (ii) in Lemma 2.3.3, and items (ii) and (iii) in Lemma 3.2.5 establish items (i) and (ii). Observe that item (iii) in Lemma 2.3.3 and item (iii) in Lemma 2.3.5 demonstrate that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) &= (\mathcal{R}_\tau^{\mathbf{N}}(0 \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\xi}))) (x) + f(\xi) \\ &= 0((\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{1,-\xi}))(x)) + f(\xi) = f(\xi) \end{aligned} \quad (3.80)$$

(cf. Definitions 1.2.4 and 1.3.4). This proves item (iii). Note that (3.80), the fact that $\xi \in [a, b]$, and the assumption that for all $x, y \in [a, b]$ it holds that

$$|f(x) - f(y)| \leq L|x - y| \quad (3.81)$$

show that for all $x \in [a, b]$ it holds that

$$|(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| = |f(\xi) - f(x)| \leq L|x - \xi| \leq L \max\{\xi - a, b - \xi\}. \quad (3.82)$$

This establishes item (iv). Observe that (1.55) and item (i) ensure that

$$\mathcal{P}(\mathbf{F}) = 1(1 + 1) + 1(1 + 1) = 4. \quad (3.83)$$

This proves item (v). The proof of Lemma 3.3.2 is thus complete. \square

Corollary 3.3.3 (Explicit ANN approximations with prescribed error tolerances).

Let $\varepsilon \in (0, \infty)$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $K \in \mathbb{N}_0 \cap \left(\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1\right)$, $\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that $\mathbf{r}_k = a + \frac{k(b-a)}{\max\{K, 1\}}$, let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for

all $x, y \in [a, b]$ that

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.84)$$

and let $\mathbf{F} \in \mathbf{N}$ satisfy

$$\mathbf{F} = \mathbf{A}_{1, f(x_0)} \bullet \left(\bigoplus_{k=0}^K \left(\left(\frac{K(f(x_{\min\{k+1, K\}}) - 2f(x_k) + f(x_{\max\{k-1, 0\}}))}{(b-a)} \right) \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1, -x_k}) \right) \right) \quad (3.85)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.2). Then

(i) it holds that $\mathcal{D}(\mathbf{F}) = (1, K + 1, 1)$,

(ii) it holds that $\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,

(iii) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,

(iv) it holds that $\sup_{x \in [a, b]} |(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \frac{L(b-a)}{\max\{K, 1\}} \leq \varepsilon$, and

(v) it holds that $\mathcal{P}(\mathbf{F}) = 3K + 4 \leq 3L(b-a)\varepsilon^{-1} + 7$

(cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Proof of Corollary 3.3.3. Note that the assumption that $K \in \mathbb{N}_0 \cap \left[\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1 \right)$ implies that

$$\frac{L(b-a)}{\max\{K, 1\}} \leq \varepsilon. \quad (3.86)$$

This, items (i), (iii), and (iv) in Proposition 3.3.1, and items (i), (ii), (iii), and (iv) in Lemma 3.3.2 establish items (i), (ii), (iii), and (iv). Observe that item (v) in Proposition 3.3.1, item (v) in Lemma 3.3.2, and the fact that

$$K \leq 1 + \frac{L(b-a)}{\varepsilon}, \quad (3.87)$$

demonstrate that

$$\mathcal{P}(\mathbf{F}) = 3K + 4 \leq \frac{3L(b-a)}{\varepsilon} + 7. \quad (3.88)$$

This proves item (v). The proof of Corollary 3.3.3 is thus complete. \square

3.3.2 Convergence rates for the approximation error

Definition 3.3.4 (Quasi vector norms). We denote by $\|\cdot\|_p: (\bigcup_{d=1}^{\infty} \mathbb{R}^d) \rightarrow \mathbb{R}$, $p \in (0, \infty]$, the functions which satisfy for all $p \in (0, \infty)$, $d \in \mathbb{N}$, $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ that

$$\|\theta\|_p = \left[\sum_{i=1}^d |\theta_i|^p \right]^{1/p} \quad \text{and} \quad \|\theta\|_{\infty} = \max_{i \in \{1, 2, \dots, d\}} |\theta_i|. \quad (3.89)$$

Corollary 3.3.5 (Implicit one-dimensional ANN approximations with prescribed error tolerances and explicit parameter bounds). Let $\varepsilon \in (0, \infty)$, $L \in [0, \infty)$, $a \in \mathbb{R}$, $b \in [a, \infty)$ and let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.90)$$

Then there exists $\mathbf{F} \in \mathbf{N}$ such that

- (i) it holds that $\mathcal{R}_{\tau}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,
 - (ii) it holds that $\mathcal{H}(\mathbf{F}) = 1$,
 - (iii) it holds that $\mathbb{D}_1(\mathbf{F}) \leq L(b - a)\varepsilon^{-1} + 2$,
 - (iv) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}_{\tau}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\tau}^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,
 - (v) it holds that $\sup_{x \in [a, b]} |(\mathcal{R}_{\tau}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$,
 - (vi) it holds that $\mathcal{P}(\mathbf{F}) = 3(\mathbb{D}_1(\mathbf{F})) + 1 \leq 3L(b - a)\varepsilon^{-1} + 7$, and
 - (vii) it holds that $\|\mathcal{T}(\mathbf{F})\|_{\infty} \leq \max\{1, |a|, |b|, 2L, |f(a)|\}$
- (cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 3.3.4).

Proof of Corollary 3.3.5. Throughout this proof, assume without loss of generality that $a < b$, let $K \in \mathbb{N}_0 \cap \left[\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1 \right)$, $\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K \in [a, b]$, $c_0, c_1, \dots, c_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that

$$\mathfrak{x}_k = a + \frac{k(b-a)}{\max\{K, 1\}} \quad \text{and} \quad c_k = \frac{K(f(\mathfrak{x}_{\min\{k+1, K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1, 0\}}))}{(b-a)}, \quad (3.91)$$

and let $\mathbf{F} \in \mathbf{N}$ satisfy

$$\mathbf{F} = \mathbf{A}_{1, f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^K (c_k \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1, -\mathfrak{x}_k})) \right) \quad (3.92)$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.2). Note that Corollary 3.3.3 shows that

- (I) it holds that $\mathcal{D}(\mathbf{F}) = (1, K + 1, 1)$,
- (II) it holds that $\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,
- (III) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,
- (IV) it holds that $\sup_{x \in [a, b]} |(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$, and
- (V) it holds that $\mathcal{P}(\mathbf{F}) = 3K + 4$

(cf. Definitions 1.2.4 and 1.3.4). This establishes items (i), (iv), and (v). Observe that item (I) and the fact that

$$K \leq 1 + \frac{L(b - a)}{\varepsilon} \quad (3.93)$$

prove items (ii) and (iii). Note that item (ii) and items (I) and (V) ensure that

$$\mathcal{P}(\mathbf{F}) = 3K + 4 = 3(K + 1) + 1 = 3(\mathbb{D}_1(\mathbf{F})) + 1 \leq \frac{3L(b - a)}{\varepsilon} + 7. \quad (3.94)$$

This proves item (vi). Observe that Lemma 3.2.5 implies that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$c_k \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1, -\mathbf{r}_k}) = ((1, -\mathbf{r}_k), (c_k, 0)). \quad (3.95)$$

Combining this with (2.120), (2.111), (2.102), and (2.2) demonstrates that

$$\mathbf{F} = \left(\left(\left(\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} -\mathbf{r}_0 \\ -\mathbf{r}_1 \\ \vdots \\ -\mathbf{r}_K \end{pmatrix} \right), ((c_0 \ c_1 \ \cdots \ c_K), f(\mathbf{r}_0)) \right) \in (\mathbb{R}^{(K+1) \times 1} \times \mathbb{R}^{K+1}) \times (\mathbb{R}^{1 \times (K+1)} \times \mathbb{R}). \quad (3.96)$$

Lemma 1.3.9 hence shows that

$$\|\mathcal{T}(\mathbf{F})\|_\infty = \max\{|\mathbf{r}_0|, |\mathbf{r}_1|, \dots, |\mathbf{r}_K|, |c_0|, |c_1|, \dots, |c_K|, |f(\mathbf{r}_0)|, 1\} \quad (3.97)$$

(cf. Definitions 1.3.6 and 3.3.4). Next note that the assumption that for all $x, y \in [a, b]$ it holds that

$$|f(x) - f(y)| \leq L|x - y| \quad (3.98)$$

and the fact that for all $k \in \mathbb{N} \cap (0, K + 1)$ it holds that

$$\mathbf{r}_k - \mathbf{r}_{k-1} = \frac{(b - a)}{\max\{K, 1\}} \quad (3.99)$$

establish that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$\begin{aligned}
 |c_k| &\leq \frac{K(|f(\mathbf{r}_{\min\{k+1, K\}}) - f(\mathbf{r}_k)| + |f(\mathbf{r}_{\max\{k-1, 0\}}) - f(\mathbf{r}_k)|)}{(b-a)} \\
 &\leq \frac{KL(|\mathbf{r}_{\min\{k+1, K\}} - \mathbf{r}_k| + |\mathbf{r}_{\max\{k-1, 0\}} - \mathbf{r}_k|)}{(b-a)} \\
 &\leq \frac{2KL(b-a)[\max\{K, 1\}]^{-1}}{(b-a)} \leq 2L.
 \end{aligned} \tag{3.100}$$

This and (3.97) prove item (vii). The proof of Corollary 3.3.5 is thus complete. \square

Corollary 3.3.6 (Implicit one-dimensional ANN approximations with prescribed error tolerances and asymptotic parameter bounds). *Let $L, a \in \mathbb{R}$, $b \in [a, \infty)$ and let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|. \tag{3.101}$$

Then there exist $\mathfrak{C} \in \mathbb{R}$ such that for all $\varepsilon \in (0, 1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}_{\dagger}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [a, b]} |(\mathcal{R}_{\dagger}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \mathcal{H}(\mathbf{F}) = 1, \tag{3.102}$$

$$\|\mathcal{T}(\mathbf{F})\|_{\infty} \leq \max\{1, |a|, |b|, 2L, |f(a)|\}, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-1} \tag{3.103}$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 3.3.4).

Proof of Corollary 3.3.6. Throughout this proof, assume without loss of generality that $a < b$ and let

$$\mathfrak{C} = 3L(b-a) + 7. \tag{3.104}$$

Observe that the assumption that $a < b$ ensures that $L \geq 0$. Next note that (3.104) implies that for all $\varepsilon \in (0, 1]$ it holds that

$$3L(b-a)\varepsilon^{-1} + 7 \leq 3L(b-a)\varepsilon^{-1} + 7\varepsilon^{-1} = \mathfrak{C}\varepsilon^{-1}. \tag{3.105}$$

This and Corollary 3.3.5 demonstrate that for all $\varepsilon \in (0, 1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}_{\dagger}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [a, b]} |(\mathcal{R}_{\dagger}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \mathcal{H}(\mathbf{F}) = 1, \tag{3.106}$$

$$\|\mathcal{T}(\mathbf{F})\|_{\infty} \leq \max\{1, |a|, |b|, 2L, |f(a)|\}, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq 3L(b-a)\varepsilon^{-1} + 7 \leq \mathfrak{C}\varepsilon^{-1} \tag{3.107}$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 3.3.4). The proof of Corollary 3.3.6 is thus complete. \square

Corollary 3.3.7 (Implicit one-dimensional ANN approximations with prescribed error tolerances and asymptotic parameter bounds). *Let $L, a \in \mathbb{R}$, $b \in [a, \infty)$ and let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.108)$$

Then there exist $\mathfrak{C} \in \mathbb{R}$ such that for all $\varepsilon \in (0, 1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [a, b]} |(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-1} \quad (3.109)$$

(cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Proof of Corollary 3.3.7. Observe that Corollary 3.3.6 establishes (3.109). The proof of Corollary 3.3.7 is thus complete. \square

Exercise 3.3.1. Let $f: [-2, 3] \rightarrow \mathbb{R}$ satisfy for all $x \in [-2, 3]$ that

$$f(x) = x^2 + 2 \sin(x). \quad (3.110)$$

Prove or disprove the following statement: There exist $c \in \mathbb{R}$ and $\mathbf{F} = (\mathbf{F}_\varepsilon)_{\varepsilon \in (0, 1]}: (0, 1] \rightarrow \mathbf{N}$ such that for all $\varepsilon \in (0, 1]$ it holds that

$$\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}_\varepsilon) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [-2, 3]} |(\mathcal{R}_\tau^{\mathbf{N}}(\mathbf{F}_\varepsilon))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}_\varepsilon) \leq c\varepsilon^{-1} \quad (3.111)$$

(cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Exercise 3.3.2. Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that $\mathcal{P}(\Phi) \leq 10$ and

$$\sup_{x \in [0, 10]} |\sqrt{x} - (\mathcal{R}_\tau^{\mathbf{N}}(\Phi))(x)| \leq \frac{1}{4} \quad (3.112)$$

(cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Part III
Optimization

Chapter 4

Optimization through gradient flow (GF) trajectories

In Chapters 5 and 6 below we study deterministic and stochastic gradient descent (SGD)-type optimization methods from the literature. Such methods are widely used in machine learning problems to approximately minimize suitable objective functions. The SGD-type optimization methods in Chapter 6 can be viewed as suitable Monte Carlo approximations of the deterministic GD-type optimization methods in Chapter 5 and the deterministic GD-type optimization methods in Chapter 5 can, roughly speaking, be viewed as time-discrete approximations of solutions of suitable GF ODEs. To develop intuitions for GD-type optimization methods and for some of the tools which we employ to analyze such methods, we study in this chapter such GF ODEs. In particular, we show in this chapter how such GF ODEs can be used to approximately solve appropriate optimization problems.

4.1 Introductory comments for the training of ANNs

Key components of deep supervised learning algorithms are typically deep ANNs and also suitable *gradient based optimization methods*. In Parts I and II we have introduced and studied different types of ANNs while in Part III we introduce and study gradient based optimization methods. In this section we briefly outline the main ideas behind gradient based optimization methods and sketch how such methods arise within deep supervised learning algorithms. To do this, we now recall the deep supervised learning framework from the [introduction](#).

Specifically, let $d, M \in \mathbb{N}$, $\mathcal{E} \in C(\mathbb{R}^d, \mathbb{R})$, $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$ satisfy for all $m \in \{1, 2, \dots, M\}$ that

$$y_m = \mathcal{E}(x_m) \tag{4.1}$$

and let $\mathfrak{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $\phi \in C(\mathbb{R}^d, \mathbb{R})$ that

$$\mathfrak{L}(\phi) = \frac{1}{M} \left[\sum_{m=1}^M |\phi(\mathbf{x}_m) - \mathbf{y}_m|^2 \right]. \quad (4.2)$$

As in the [introduction](#) we think of $M \in \mathbb{N}$ as the number of available known input-output data pairs, we think of $d \in \mathbb{N}$ as the dimension of the input data, we think of $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ as an unknown function which relates input and output data through (4.1), we think of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{M+1} \in \mathbb{R}^d$ as the available known input data, we think of $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}$ as the available known output data, and we have that the function $\mathfrak{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$ in (4.2) is the objective function (the function we want to minimize) in the optimization problem associated to the considered learning problem (cf. (3) in the [introduction](#)). In particular, observe that

$$\mathfrak{L}(\mathcal{E}) = 0 \quad (4.3)$$

and we are trying to approximate the function \mathcal{E} by computing an approximate minimizer of the function $\mathfrak{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$. In order to make this optimization problem amenable to numerical computations, we consider a spatially discretized version of the optimization problem associated to (4.2) by employing parametrizations of ANNs (cf. (7) in the [introduction](#)).

More formally, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \dots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, and consider the parametrization function

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d} \in C(\mathbb{R}^d, \mathbb{R}) \quad (4.4)$$

(cf. Definitions 1.1.3 and 1.2.1). Note that h is the number of hidden layers of the ANNs in (4.4), note for every $i \in \{1, 2, \dots, h\}$ that $l_i \in \mathbb{N}$ is the number of neurons in the i -th hidden layer of the ANNs in (4.4), and note that \mathfrak{d} is the number of real parameters used to describe the ANNs in (4.4). Observe that for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ we have that the function

$$\mathbb{R}^d \ni x \mapsto \mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d}(x) \in \mathbb{R} \quad (4.5)$$

in (4.4) is nothing else than the realization function associated to an ANN where before each hidden layer a multidimensional version of the activation function $a: \mathbb{R} \rightarrow \mathbb{R}$ is applied. We restrict ourselves in this section to a differentiable activation function as this differentiability property allows us to consider gradients (cf. (4.7), (4.8) below for details).

We now discretize the optimization problem in (4.2) as the problem of computing approximate minimizers of the function $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ which satisfies for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M |(\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d})(\mathbf{x}_m) - \mathbf{y}_m|^2 \right] \quad (4.6)$$

and this resulting optimization problem is now accessible to numerical computations. Specifically, deep learning algorithms solve optimization problems of the type (4.6) by means of *gradient based optimization methods*. Loosely speaking, gradient based optimization methods aim to minimize the considered objective function (such as (4.6) above) by performing successive steps based on the direction of the negative gradient of the objective function. One of the simplest gradient based optimization method is the plain-vanilla GD optimization method which performs successive steps in the direction of the negative gradient. We now sketch the GD optimization method applied to (4.6). Let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, and let $\theta = (\theta_n)_{n \in \mathbb{N}_0} : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\theta_0 = \xi \quad \text{and} \quad \theta_n = \theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\theta_{n-1}). \quad (4.7)$$

The process $(\theta_n)_{n \in \mathbb{N}_0}$ is the GD process for the minimization problem associated to (4.6) with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ (see Definition 5.1.1 below for the precise definition).

This plain-vanilla GD optimization method and related GD-type optimization methods can be regarded as discretizations of solutions of GF ODEs. In the context of the minimization problem in (4.6) such solutions of GF ODEs can be described as follows. Let $\Theta = (\Theta_t)_{t \in [0, \infty)} : [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a continuously differentiable function which satisfies for all $t \in [0, \infty)$ that

$$\Theta_0 = \xi \quad \text{and} \quad \dot{\Theta}_t = \frac{\partial}{\partial t} \Theta_t = -(\nabla \mathcal{L})(\Theta_t). \quad (4.8)$$

The process $(\Theta_t)_{t \in [0, \infty)}$ is the solution of the GF ODE corresponding to the minimization problem associated to (4.6) with initial value ξ .

In Chapter 5 below we introduce and study deterministic GD-type optimization methods such as the GD optimization method in (4.7). To develop intuitions for GD-type optimization methods and for some of the tools which we employ to analyze such GD-type optimization methods, we study in the remainder of this chapter GF ODEs such as (4.8) above. In deep learning algorithms usually stochastic variants of GD-type optimization methods are employed to solve optimization problems of the form (4.6). Such SGD-type optimization methods can be viewed as suitable Monte Carlo approximations of deterministic GD-type methods and in Chapter 6 below we treat such SGD-type optimization methods.

Definition 4.1.1 (GF trajectories). *Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $\mathcal{L} : \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function, and let $\mathcal{G} : \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a $\mathcal{B}(\mathbb{R}^{\mathfrak{d}})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable function which satisfies for all open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ and $\theta \in U$ with $\mathcal{L}|_U \in C^1(U, \mathbb{R})$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta). \quad (4.9)$$

Then we say that Θ is a GF trajectory for the objective function \mathcal{L} with generalized gradient \mathcal{G} and initial value ξ if $\Theta : [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ is a function from $[0, \infty)$ to $\mathbb{R}^{\mathfrak{d}}$

which satisfies for all $t \in [0, \infty)$ that

$$\Theta_t = \xi - \int_0^t \mathcal{G}(\Theta_s) ds. \quad (4.10)$$

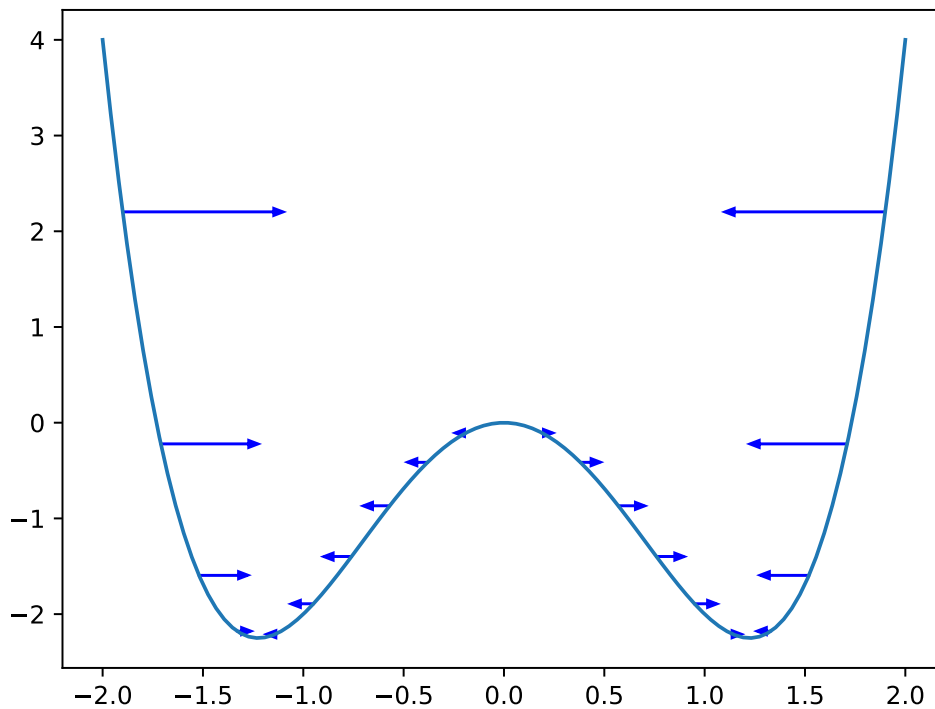


Figure 4.1 ([plots/gradient_plot1.pdf](#)): Illustration of negative gradients in a one-dimensional example. The plot shows the graph of the function $[-2, 2] \ni x \mapsto x^4 - 3x^2 \in \mathbb{R}$ with the value of the negative gradient at several points indicated by horizontal arrows.

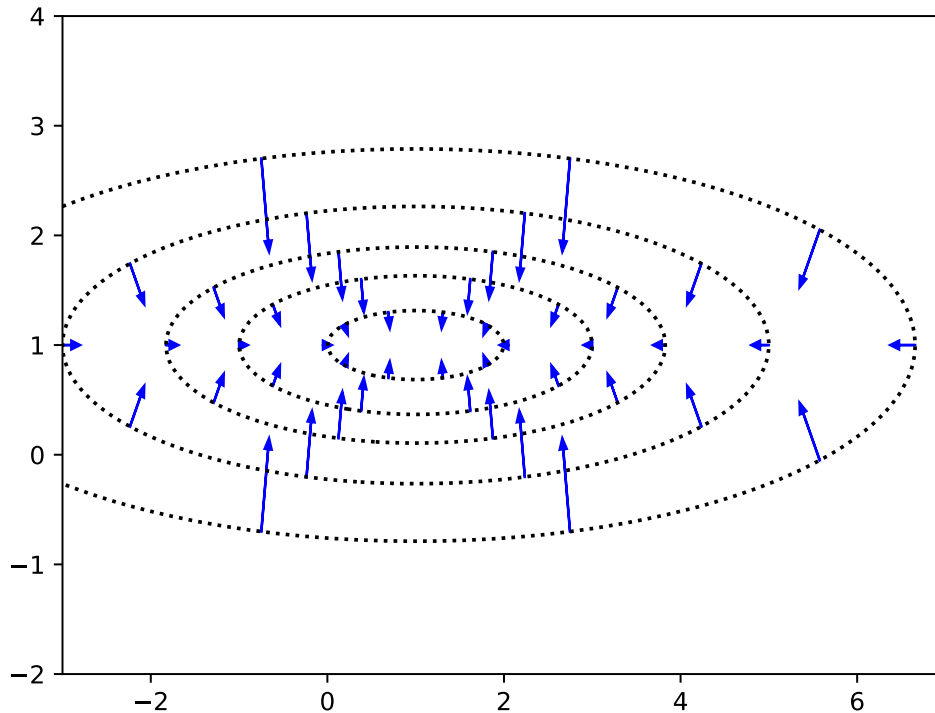


Figure 4.2 ([plots/gradient_plot2.pdf](#)): Illustration of negative gradients in a two-dimensional example. The plot shows contour lines of the function $\mathbb{R}^2 \ni (x, y) \mapsto \frac{1}{2}|x - 1|^2 + 5|y - 1|^2 \in \mathbb{R}$ with arrows indicating the direction and magnitude of the negative gradient at several points along these contour lines.

4.2 Loss functions

4.2.1 Absolute error loss

Definition 4.2.1. Let $d \in \mathbb{N}$ and let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ be a norm. We say that \mathbf{L} is the ℓ^1 -error loss function based on $\|\cdot\|$ if $\mathbf{L}: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the function from $\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R} which satisfies for all $x, y \in \mathbb{R}^d$ that

$$\mathbf{L}(x, y) = \|x - y\|. \quad (4.11)$$

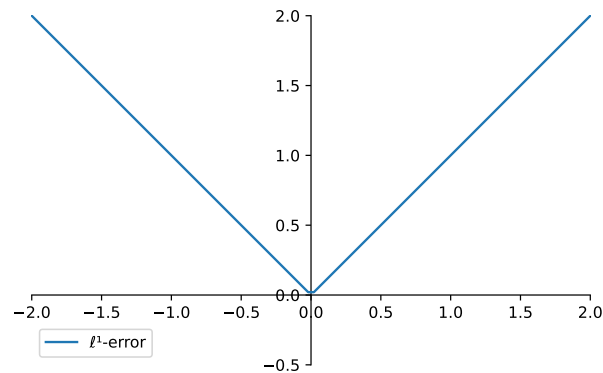


Figure 4.3 ([plots/l1loss.pdf](#)): A plot of the function $\mathbb{R} \ni x \mapsto \mathbf{L}(x, 0) \in [0, \infty)$ where \mathbf{L} is the ℓ^1 -error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$ (cf. Definition 4.2.1).

4.2.2 Mean squared error loss

Definition 4.2.2. Let $d \in \mathbb{N}$ and let $\|\cdot\| : \mathbb{R}^d \rightarrow [0, \infty)$ be a norm. We say that \mathbf{L} is the mean squared error loss function based on $\|\cdot\|$ if $\mathbf{L} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the function from $\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R} which satisfies for all $x, y \in \mathbb{R}^d$ that

$$\mathbf{L}(x, y) = \|x - y\|^2. \quad (4.12)$$

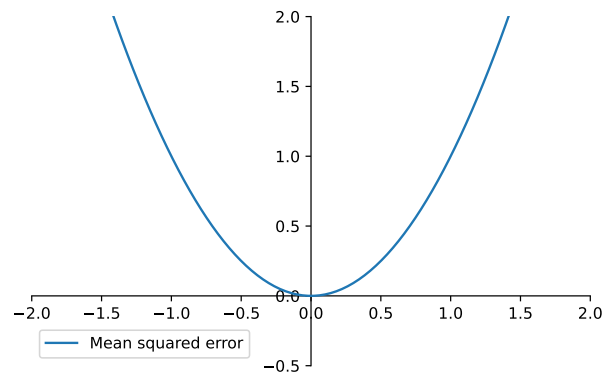


Figure 4.4 ([plots/mseloss.pdf](#)): A plot of the function $\mathbb{R} \ni x \mapsto \mathbf{L}(x, 0) \in [0, \infty)$ where \mathbf{L} is the mean squared error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$ (cf. Definition 4.2.2).

Lemma 4.2.3. Let $d \in \mathbb{N}$ and let \mathbf{L} be the mean squared error loss function based on $\mathbb{R}^d \ni x \mapsto \|x\|_2 \in [0, \infty)$ (cf. Definitions 3.3.4 and 4.2.2). Then

(i) it holds that $\mathbf{L} \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$

(ii) it holds for all $x, y, u, v \in \mathbb{R}^d$ that

$$\mathbf{L}(u, v) = \mathbf{L}(x, y) + \mathbf{L}'(x, y)(u - x, v - y) + \frac{1}{2} \mathbf{L}^{(2)}(x, y)((u - x, v - y), (u - x, v - y)). \quad (4.13)$$

4.3 GF optimization in the training of ANNs

Example 4.3.1. Let $d, L, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_L \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d + 1) + [\sum_{k=2}^L l_k(l_{k-1} + 1)]$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M \in \mathbb{R}^d$, $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}^{l_L}$, let $\mathbf{L}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be the mean squared error loss function based on $\mathbb{R}^d \ni x \mapsto \|x\|_2 \in [0, \infty)$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}((\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_L}, \text{id}_{\mathbb{R}^{l_L}}})^{\theta, d})(\mathbf{x}_m), \mathbf{y}_m) \right], \quad (4.14)$$

let $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta: [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $t \in [0, \infty)$ that

$$\Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds \quad (4.15)$$

(cf. Definitions 1.1.3, 1.2.1, 3.3.4, and 4.2.2, and Lemma 4.2.3). Then Θ is a GF trajectory for the objective function \mathcal{L} with initial value ξ (cf. Definition 4.1.1).

Proof for Example 4.3.1. Note that (4.9), (4.10), and (4.15) demonstrate that Θ is a GF trajectory for the objective function \mathcal{L} with initial value ξ (cf. Definition 4.1.1). The proof for Example 4.3.1 is thus complete. \square

4.4 Lyapunov-type functions for GFs

4.4.1 Gronwall differential inequalities

The following lemma, Lemma 4.4.1 below, is referred to as a Gronwall inequality in the literature. Gronwall inequalities are powerful tools to study dynamical systems and,

especially, solutions of ODEs.

Lemma 4.4.1 (Gronwall inequality). *Let $T \in (0, \infty)$, $\alpha \in \mathbb{R}$, $\epsilon \in C^1([0, T], \mathbb{R})$, $\beta \in C([0, T], \mathbb{R})$ satisfy for all $t \in [0, T]$ that*

$$\epsilon'(t) \leq \alpha\epsilon(t) + \beta(t). \quad (4.16)$$

Then it holds for all $t \in [0, T]$ that

$$\epsilon(t) \leq e^{\alpha t}\epsilon(0) + \int_0^t e^{\alpha(t-s)}\beta(s) \, ds. \quad (4.17)$$

Proof of Lemma 4.4.1. Throughout this proof, let $v: [0, T] \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$ that

$$v(t) = e^{\alpha t} \left[\int_0^t e^{-\alpha s} \beta(s) \, ds \right] \quad (4.18)$$

and let $u: [0, T] \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$ that

$$u(t) = [\epsilon(t) - v(t)]e^{-\alpha t}. \quad (4.19)$$

Observe that the product rule and the fundamental theorem of calculus demonstrate that for all $t \in [0, T]$ it holds that $v \in C^1([0, T], \mathbb{R})$ and

$$v'(t) = \left[\int_0^t \alpha e^{\alpha(t-s)} \beta(s) \, ds \right] + \beta(t) = \alpha \left[\int_0^t e^{\alpha(t-s)} \beta(s) \, ds \right] + \beta(t) = \alpha v(t) + \beta(t). \quad (4.20)$$

The assumption that $\epsilon \in C^1([0, T], \mathbb{R})$ and the product rule therefore ensure that for all $t \in [0, T]$ it holds that $u \in C^1([0, T], \mathbb{R})$ and

$$\begin{aligned} u'(t) &= [\epsilon'(t) - v'(t)]e^{-\alpha t} - [\epsilon(t) - v(t)]\alpha e^{-\alpha t} \\ &= [\epsilon'(t) - v'(t) - \alpha\epsilon(t) + \alpha v(t)]e^{-\alpha t} \\ &= [\epsilon'(t) - \alpha v(t) - \beta(t) - \alpha\epsilon(t) + \alpha v(t)]e^{-\alpha t} \\ &= [\epsilon'(t) - \beta(t) - \alpha\epsilon(t)]e^{-\alpha t}. \end{aligned} \quad (4.21)$$

Combining this with the assumption that for all $t \in [0, T]$ it holds that $\epsilon'(t) \leq \alpha\epsilon(t) + \beta(t)$ proves that for all $t \in [0, T]$ it holds that

$$u'(t) \leq [\alpha\epsilon(t) + \beta(t) - \beta(t) - \alpha\epsilon(t)]e^{-\alpha t} = 0. \quad (4.22)$$

This and the fundamental theorem of calculus imply that for all $t \in [0, T]$ it holds that

$$u(t) = u(0) + \int_0^t u'(s) \, ds \leq u(0) + \int_0^t 0 \, ds = u(0) = \epsilon(0). \quad (4.23)$$

Combining this, (4.18), and (4.19) shows that for all $t \in [0, T]$ it holds that

$$\epsilon(t) = e^{\alpha t}u(t) + v(t) \leq e^{\alpha t}\epsilon(0) + v(t) \leq e^{\alpha t}\epsilon(0) + \int_0^t e^{\alpha(t-s)}\beta(s) ds. \quad (4.24)$$

The proof of Lemma 4.4.1 is thus complete. \square

4.4.2 Lyapunov-type functions for ODEs

Proposition 4.4.2 (Lyapunov-type functions for ODEs). *Let $\mathfrak{d} \in \mathbb{N}$, $T \in (0, \infty)$, $\alpha \in \mathbb{R}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\beta \in C(O, \mathbb{R})$, $\mathcal{G} \in C(O, \mathbb{R}^{\mathfrak{d}})$, $V \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$V'(\theta)\mathcal{G}(\theta) = \langle (\nabla V)(\theta), \mathcal{G}(\theta) \rangle \leq \alpha V(\theta) + \beta(\theta), \quad (4.25)$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds$. Then it holds for all $t \in [0, T]$ that

$$V(\Theta_t) \leq e^{\alpha t}V(\Theta_0) + \int_0^t e^{\alpha(t-s)}\beta(\Theta_s) ds. \quad (4.26)$$

Proof of Proposition 4.4.2. Throughout this proof, let $\epsilon, b \in C([0, T], \mathbb{R})$ satisfy for all $t \in [0, T]$ that

$$\epsilon(t) = V(\Theta_t) \quad \text{and} \quad b(t) = \beta(\Theta_t). \quad (4.27)$$

Note that (4.25), (4.27), the fundamental theorem of calculus, and the chain rule ensure that for all $t \in [0, T]$ it holds that

$$\epsilon'(t) = \frac{d}{dt}(V(\Theta_t)) = V'(\Theta_t)(\dot{\Theta}_t) = V'(\Theta_t)\mathcal{G}(\Theta_t) \leq \alpha V(\Theta_t) + \beta(\Theta_t) = \alpha\epsilon(t) + b(t). \quad (4.28)$$

Lemma 4.4.1 and (4.27) hence demonstrate that for all $t \in [0, T]$ it holds that

$$V(\Theta_t) = \epsilon(t) \leq \epsilon(0)e^{\alpha t} + \int_0^t e^{\alpha(t-s)}b(s) ds = V(\Theta_0)e^{\alpha t} + \int_0^t e^{\alpha(t-s)}\beta(\Theta_s) ds. \quad (4.29)$$

The proof of Proposition 4.4.2 is thus complete. \square

Corollary 4.4.3. *Let $\mathfrak{d} \in \mathbb{N}$, $T \in (0, \infty)$, $\alpha \in \mathbb{R}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\mathcal{G} \in C(O, \mathbb{R}^{\mathfrak{d}})$, $V \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$V'(\theta)\mathcal{G}(\theta) = \langle (\nabla V)(\theta), \mathcal{G}(\theta) \rangle \leq \alpha V(\theta), \quad (4.30)$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds$. Then it

holds for all $t \in [0, T]$ that

$$V(\Theta_t) \leq e^{\alpha t} V(\Theta_0). \quad (4.31)$$

Proof of Corollary 4.4.3. Observe that Proposition 4.4.2 and (4.30) establish (4.31). The proof of Corollary 4.4.3 is thus complete. \square

Corollary 4.4.4 (On quadratic Lyapunov-type functions and coercivity-type conditions). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$, $T \in (0, \infty)$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\mathcal{L} \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2, \quad (4.32)$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$. Then it holds for all $t \in [0, T]$ that

$$\|\Theta_t - \vartheta\|_2 \leq e^{-ct} \|\Theta_0 - \vartheta\|_2. \quad (4.33)$$

4.4.3 Sufficient and necessary conditions for local minimum points

Lemma 4.4.5. *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\vartheta \in O$, let $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function, assume that \mathcal{L} is differentiable at ϑ , and assume that $(\nabla \mathcal{L})(\vartheta) \neq 0$. Then there exists $\theta \in O$ such that $\mathcal{L}(\theta) < \mathcal{L}(\vartheta)$.*

Proof of Lemma 4.4.5. Throughout this proof, let $v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}$ satisfy $v = -(\nabla \mathcal{L})(\vartheta)$, let $\delta \in (0, \infty)$ satisfy for all $t \in (-\delta, \delta)$ that

$$\vartheta + tv = \vartheta - t(\nabla \mathcal{L})(\vartheta) \in O, \quad (4.34)$$

and let $L: (-\delta, \delta) \rightarrow \mathbb{R}$ satisfy for all $t \in (-\delta, \delta)$ that

$$L(t) = \mathcal{L}(\vartheta + tv). \quad (4.35)$$

Note that for all $t \in (0, \delta)$ it holds that

$$\begin{aligned} & \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| = \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] + \|(\nabla \mathcal{L})(\vartheta)\|_2^2 \right| \\ & = \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] + \langle (\nabla \mathcal{L})(\vartheta), (\nabla \mathcal{L})(\vartheta) \rangle \right| \\ & = \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] - \langle (\nabla \mathcal{L})(\vartheta), v \rangle \right|. \end{aligned} \quad (4.36)$$

Therefore, we obtain that for all $t \in (0, \delta)$ it holds that

$$\begin{aligned} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| &= \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] - \mathcal{L}'(\vartheta)v \right| \\ &= \left| \frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta) - \mathcal{L}'(\vartheta)tv}{t} \right| = \frac{|\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta) - \mathcal{L}'(\vartheta)tv|}{t}. \end{aligned} \quad (4.37)$$

The assumption that \mathcal{L} is differentiable at ϑ hence demonstrates that

$$\limsup_{t \searrow 0} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| = 0. \quad (4.38)$$

The fact that $\|v\|_2^2 > 0$ therefore demonstrates that there exists $t \in (0, \delta)$ such that

$$\left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| < \frac{\|v\|_2^2}{2}. \quad (4.39)$$

The triangle inequality and the fact that $\|v\|_2^2 > 0$ hence prove that

$$\begin{aligned} \frac{L(t) - L(0)}{t} &= \left[\frac{L(t) - L(0)}{t} + \|v\|_2^2 \right] - \|v\|_2^2 \leq \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| - \|v\|_2^2 \\ &< \frac{\|v\|_2^2}{2} - \|v\|_2^2 = -\frac{\|v\|_2^2}{2} < 0. \end{aligned} \quad (4.40)$$

This ensures that

$$\mathcal{L}(\vartheta + tv) = L(t) < L(0) = \mathcal{L}(\vartheta). \quad (4.41)$$

The proof of Lemma 4.4.5 is thus complete. \square

Lemma 4.4.6 (A necessary condition for a local minimum point). *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\vartheta \in O$, let $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function, assume that \mathcal{L} is differentiable at ϑ , and assume*

$$\mathcal{L}(\vartheta) = \inf_{\theta \in O} \mathcal{L}(\theta). \quad (4.42)$$

Then $(\nabla \mathcal{L})(\vartheta) = 0$.

Proof of Lemma 4.4.6. We prove Lemma 4.4.6 by contradiction. We thus assume that $(\nabla \mathcal{L})(\vartheta) \neq 0$. Lemma 4.4.5 then implies that there exists $\theta \in O$ such that $\mathcal{L}(\theta) < \mathcal{L}(\vartheta)$. Combining this with (4.42) shows that

$$\mathcal{L}(\theta) < \mathcal{L}(\vartheta) = \inf_{w \in O} \mathcal{L}(w) \leq \mathcal{L}(\theta). \quad (4.43)$$

The proof of Lemma 4.4.6 is thus complete. \square

Lemma 4.4.7 (A sufficient condition for a local minimum point). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad (4.44)$$

Then

- (i) *it holds for all $\theta \in \mathbb{B}$ that $\mathcal{L}(\theta) - \mathcal{L}(\vartheta) \geq \frac{c}{2} \|\theta - \vartheta\|_2^2$,*
- (ii) *it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$, and*
- (iii) *it holds that $(\nabla \mathcal{L})(\vartheta) = 0$.*

Proof of Lemma 4.4.7. Throughout this proof, let B be the set given by

$$B = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 < r\}. \quad (4.45)$$

Note that (4.44) implies that for all $v \in \mathbb{R}^{\mathfrak{d}}$ with $\|v\|_2 \leq r$ it holds that

$$\langle (\nabla \mathcal{L})(\vartheta + v), v \rangle \geq c \|v\|_2^2. \quad (4.46)$$

The fundamental theorem of calculus hence demonstrates that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \mathcal{L}(\theta) - \mathcal{L}(\vartheta) &= [\mathcal{L}(\vartheta + t(\theta - \vartheta))]_{t=0}^{t=1} \\ &= \int_0^1 \mathcal{L}'(\vartheta + t(\theta - \vartheta))(\theta - \vartheta) dt \\ &= \int_0^1 \langle (\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta)), t(\theta - \vartheta) \rangle \frac{1}{t} dt \\ &\geq \int_0^1 c \|t(\theta - \vartheta)\|_2^2 \frac{1}{t} dt = c \|\theta - \vartheta\|_2^2 \left[\int_0^1 t dt \right] = \frac{c}{2} \|\theta - \vartheta\|_2^2. \end{aligned} \quad (4.47)$$

This proves item (i). Next observe that (4.47) ensures that for all $\theta \in \mathbb{B} \setminus \{\vartheta\}$ it holds that

$$\mathcal{L}(\theta) \geq \mathcal{L}(\vartheta) + \frac{c}{2} \|\theta - \vartheta\|_2^2 > \mathcal{L}(\vartheta). \quad (4.48)$$

Hence, we obtain for all $\theta \in \mathbb{B} \setminus \{\vartheta\}$ that

$$\inf_{w \in \mathbb{B}} \mathcal{L}(w) = \mathcal{L}(\vartheta) < \mathcal{L}(\theta). \quad (4.49)$$

This establishes item (ii). It thus remains thus remains to prove item (iii). For this observe that item (ii) ensures that

$$\{\theta \in B : \mathcal{L}(\theta) = \inf_{w \in B} \mathcal{L}(w)\} = \{\vartheta\}. \quad (4.50)$$

Combining this, the fact that B is open, and Lemma 4.4.6 (applied with $\mathfrak{d} \curvearrowright \mathfrak{d}$, $O \curvearrowright B$, $\vartheta \curvearrowright \vartheta$, $\mathcal{L} \curvearrowright \mathcal{L}|_B$ in the notation of Lemma 4.4.6) assures that $(\nabla \mathcal{L})(\vartheta) = 0$. This establishes item (iii). The proof of Lemma 4.4.7 is thus complete. \square

4.4.4 On a linear growth condition

Lemma 4.4.8 (On a linear growth condition). *Let $\mathfrak{d} \in \mathbb{N}$, $L \in \mathbb{R}$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\|(\nabla \mathcal{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2. \quad (4.51)$$

Then it holds for all $\theta \in \mathbb{B}$ that

$$\mathcal{L}(\theta) - \mathcal{L}(\vartheta) \leq \frac{L}{2}\|\theta - \vartheta\|_2^2. \quad (4.52)$$

Proof of Lemma 4.4.8. Observe that (4.51), the Cauchy-Schwarz inequality, and the fundamental theorem of calculus ensure that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \mathcal{L}(\theta) - \mathcal{L}(\vartheta) &= [\mathcal{L}(\vartheta + t(\theta - \vartheta))]_{t=0}^{t=1} \\ &= \int_0^1 \mathcal{L}'(\vartheta + t(\theta - \vartheta))(\theta - \vartheta) dt \\ &= \int_0^1 \langle (\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta)), \theta - \vartheta \rangle dt \\ &\leq \int_0^1 \|(\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta))\|_2 \|\theta - \vartheta\|_2 dt \\ &\leq \int_0^1 L\|\vartheta + t(\theta - \vartheta) - \vartheta\|_2 \|\theta - \vartheta\|_2 dt \\ &= L\|\theta - \vartheta\|_2^2 \left[\int_0^1 t dt \right] = \frac{L}{2}\|\theta - \vartheta\|_2^2. \end{aligned} \quad (4.53)$$

The proof of Lemma 4.4.8 is thus complete. \square

4.5 Optimization through flows of ODEs

4.5.1 Approximation of local minimum points through GFs

Proposition 4.5.1 (Approximation of local minimum points through GFs). *Let $\mathfrak{d} \in \mathbb{N}$, $c, T \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2, \quad (4.54)$$

and let $\Theta \in C([0, T], \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, T]$ that $\Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$. Then

- (i) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds for all $t \in [0, T]$ that $\|\Theta_t - \vartheta\|_2 \leq e^{-ct} \|\xi - \vartheta\|_2$, and
- (iii) it holds for all $t \in [0, T]$ that

$$0 \leq \frac{c}{2} \|\Theta_t - \vartheta\|_2^2 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta). \quad (4.55)$$

Proof of Proposition 4.5.1. Throughout this proof, let $V : \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $V(\theta) = \|\theta - \vartheta\|_2^2$, let $\epsilon : [0, T] \rightarrow [0, \infty)$ satisfy for all $t \in [0, T]$ that $\epsilon(t) = \|\Theta_t - \vartheta\|_2^2 = V(\Theta_t)$, and let $\tau \in [0, T]$ be the real number given by

$$\tau = \inf(\{t \in [0, T] : \Theta_t \notin \mathbb{B}\} \cup \{T\}) = \inf(\{t \in [0, T] : \epsilon(t) > r^2\} \cup \{T\}). \quad (4.56)$$

Note that (4.54) and item (ii) in Lemma 4.4.7 establish item (i). Next observe that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $V \in C^1(\mathbb{R}^{\mathfrak{d}}, [0, \infty))$ and

$$(\nabla V)(\theta) = 2(\theta - \vartheta). \quad (4.57)$$

Moreover, observe that the fundamental theorem of calculus and the fact that $\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathcal{L})(v) \in \mathbb{R}^{\mathfrak{d}}$ and $\Theta : [0, T] \rightarrow \mathbb{R}^{\mathfrak{d}}$ are continuous functions ensure that for all $t \in [0, T]$ it holds that $\Theta \in C^1([0, T], \mathbb{R}^{\mathfrak{d}})$ and

$$\frac{d}{dt}(\Theta_t) = -(\nabla \mathcal{L})(\Theta_t). \quad (4.58)$$

Combining (4.54) and (4.57) hence demonstrates that for all $t \in [0, \tau]$ it holds that $\epsilon \in C^1([0, T], [0, \infty))$ and

$$\begin{aligned} \epsilon'(t) &= \frac{d}{dt}(V(\Theta_t)) = V'(\Theta_t)\left(\frac{d}{dt}(\Theta_t)\right) \\ &= \langle (\nabla V)(\Theta_t), \frac{d}{dt}(\Theta_t) \rangle \\ &= \langle 2(\Theta_t - \vartheta), -(\nabla \mathcal{L})(\Theta_t) \rangle \\ &= -2\langle (\Theta_t - \vartheta), (\nabla \mathcal{L})(\Theta_t) \rangle \\ &\leq -2c\|\Theta_t - \vartheta\|_2^2 = -2c\epsilon(t). \end{aligned} \quad (4.59)$$

The Gronwall inequality, for instance, in Lemma 4.4.1 therefore implies that for all $t \in [0, \tau]$ it holds that

$$\epsilon(t) \leq \epsilon(0)e^{-2ct}. \quad (4.60)$$

Hence, we obtain for all $t \in [0, \tau]$ that

$$\|\Theta_t - \vartheta\|_2 = \sqrt{\epsilon(t)} \leq \sqrt{\epsilon(0)}e^{-ct} = \|\Theta_0 - \vartheta\|_2 e^{-ct} = \|\xi - \vartheta\|_2 e^{-ct}. \quad (4.61)$$

In the next step we prove that

$$\tau > 0. \quad (4.62)$$

In our proof of (4.62) we distinguish between the case $\varepsilon(0) = 0$ and the case $\varepsilon(0) > 0$. We first prove (4.62) in the case

$$\varepsilon(0) = 0. \quad (4.63)$$

Note that (4.63), the assumption that $r \in (0, \infty]$, and the fact that $\varepsilon: [0, T] \rightarrow [0, \infty)$ is a continuous function show that

$$\tau = \inf(\{t \in [0, T]: \varepsilon(t) > r^2\} \cup \{T\}) > 0. \quad (4.64)$$

This establishes (4.62) in the case $\varepsilon(0) = 0$. In the next step we prove (4.62) in the case

$$\varepsilon(0) > 0. \quad (4.65)$$

Observe that (4.59) and the assumption that $c \in (0, \infty)$ assure that for all $t \in [0, \tau]$ with $\varepsilon(t) > 0$ it holds that

$$\varepsilon'(t) \leq -2c\varepsilon(t) < 0. \quad (4.66)$$

Combining this with (4.65) shows that

$$\varepsilon'(0) < 0. \quad (4.67)$$

The fact that $\varepsilon': [0, T] \rightarrow [0, \infty)$ is a continuous function and the assumption that $T \in (0, \infty)$ therefore demonstrate that

$$\inf(\{t \in [0, T]: \varepsilon'(t) > 0\} \cup \{T\}) > 0. \quad (4.68)$$

Next note that the fundamental theorem of calculus and the assumption that $\xi \in \mathbb{B}$ imply that for all $s \in [0, T]$ with $s < \inf(\{t \in [0, T]: \varepsilon'(t) > 0\} \cup \{T\})$ it holds that

$$\varepsilon(s) = \varepsilon(0) + \int_0^s \varepsilon'(u) du \leq \varepsilon(0) = \|\xi - \vartheta\|_2^2 \leq r^2. \quad (4.69)$$

Combining this with (4.68) proves that

$$\tau = \inf(\{s \in [0, T]: \varepsilon(s) > r^2\} \cup \{T\}) > 0. \quad (4.70)$$

This establishes (4.62) in the case $\varepsilon(0) > 0$. Note that (4.61), (4.62), and the assumption that $c \in (0, \infty)$ demonstrate that

$$\|\Theta_\tau - \vartheta\|_2 \leq \|\xi - \vartheta\|_2 e^{-c\tau} < r. \quad (4.71)$$

The fact that $\varepsilon: [0, T] \rightarrow [0, \infty)$ is a continuous function, (4.56), and (4.62) hence assure that $\tau = T$. Combining this with (4.61) proves that for all $t \in [0, T]$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq \|\xi - \vartheta\|_2 e^{-ct}. \quad (4.72)$$

This establishes item (ii). It thus remains to prove item (iii). For this observe that (4.54) and item (i) in Lemma 4.4.7 demonstrate that for all $\theta \in \mathbb{B}$ it holds that

$$0 \leq \frac{\varepsilon}{2} \|\theta - \vartheta\|_2^2 \leq \mathcal{L}(\theta) - \mathcal{L}(\vartheta). \quad (4.73)$$

Combining this and item (ii) implies that for all $t \in [0, T]$ it holds that

$$0 \leq \frac{\varepsilon}{2} \|\Theta_t - \vartheta\|_2^2 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \quad (4.74)$$

This establishes item (iii). The proof of Proposition 4.5.1 is thus complete. \square

Chapter 5

Deterministic gradient descent (GD) optimization methods

This chapter reviews and studies deterministic GD-type optimization methods.

5.1 GD optimization

In this section we review and study the classical plain-vanilla GD optimization method. A simple intuition behind the GD optimization method is the idea to solve a minimization problem by performing successive steps in direction of the steepest descents of the objective function, that is, by performing successive steps in the opposite direction of the gradients of the objective function.

A slightly different and maybe a bit more accurate perspective for the GD optimization method is to view the GD optimization method as a plain-vanilla Euler discretization of the associated GF ODE.

Definition 5.1.1 (GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ and $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all open $U \subseteq \mathbb{R}^{\mathfrak{d}}$, $\theta \in U$ with $\mathcal{L}|_U \in C^1(U, \mathbb{R}^{\mathfrak{d}})$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta). \quad (5.1)$$

Then we say that Θ is the GD process for the objective function \mathcal{L} with generalized gradient \mathcal{G} , learning rates $(\gamma_n)_{n \in \mathbb{N}}$, and initial value ξ if it holds that $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ is the function from \mathbb{N}_0 to $\mathbb{R}^{\mathfrak{d}}$ which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathcal{G}(\Theta_{n-1}). \quad (5.2)$$

Exercise 5.1.1. Let $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$ satisfy $\xi = (1, 2, 3)$, let $\mathcal{L}: \mathbb{R}^3 \rightarrow \mathbb{R}$ satisfy for all

$\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ that

$$\mathcal{L}(\theta) = 2(\theta_1)^2 + (\theta_2 + 1)^2 + (\theta_3 - 1)^2, \quad (5.3)$$

and let Θ be the GD process for the objective function \mathcal{L} with learning rates $\mathbb{N} \ni n \mapsto \frac{1}{2^n}$, and initial value ξ (cf. Definition 5.1.1). Specify Θ_1 , Θ_2 , and Θ_3 explicitly and prove that your results are correct.

Exercise 5.1.2. Let $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$ satisfy $\xi = (\xi_1, \xi_2, \xi_3) = (3, 4, 5)$, let $\mathcal{L}: \mathbb{R}^3 \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ that

$$\mathcal{L}(\theta) = (\theta_1)^2 + (\theta_2 - 1)^2 + 2(\theta_3 + 1)^2,$$

and let Θ be the GD process for the objective function \mathcal{L} with learning rates $\mathbb{N} \ni n \mapsto 1/3 \in [0, \infty)$ and initial value ξ (cf. Definition 5.1.1). Specify Θ_1 , Θ_2 , and Θ_3 explicitly and prove that your results are correct.

5.1.1 GD optimization in the training of ANNs

In the next example we apply the GD optimization method in the context of the training of ANNs in the vectorized description (see Section 1.1) with the loss function being the mean squared error loss function in Definition 4.2.2 (see Section 4.2.2).

Example 5.1.2. Let $d, h, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_h \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1} + 1)] + l_h + 1$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \left| (\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d})(x_m) - y_m \right|^2 \right], \quad (5.4)$$

let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}) \quad (5.5)$$

(cf. Definitions 1.1.3 and 1.2.1). Then Θ is the GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ .

Proof for Example 5.1.2. Observe that (5.5), (5.1), and (5.2) demonstrate that Θ is the GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ . The proof for Example 5.1.2 is thus complete. \square

5.1.2 Euler discretizations for GF ODEs

Theorem 5.1.3 (Taylor's formula). *Let $N \in \mathbb{N}$, $\alpha \in \mathbb{R}$, $\beta \in (\alpha, \infty)$, $a, b \in [\alpha, \beta]$, $f \in C^N([\alpha, \beta], \mathbb{R})$. Then*

$$f(b) = \left[\sum_{n=0}^{N-1} \frac{f^{(n)}(a)(b-a)^n}{n!} \right] + \int_0^1 \frac{f^{(N)}(a+r(b-a))(b-a)^N(1-r)^{N-1}}{(N-1)!} dr. \quad (5.6)$$

Lemma 5.1.4 (Local error of the Euler method). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathcal{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $x, y \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds, \quad \theta = \Theta_T + \gamma \mathcal{G}(\Theta_T), \quad (5.7)$$

$$\|\mathcal{G}(x)\|_2 \leq c, \quad \text{and} \quad \|\mathcal{G}'(x)y\|_2 \leq c\|y\|_2 \quad (5.8)$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2 \gamma^2. \quad (5.9)$$

Proof of Lemma 5.1.4. Note that the fundamental theorem of calculus, the hypothesis that $\mathcal{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$, and (5.7) assure that for all $t \in (0, \infty)$ it holds that $\Theta \in C^1([0, \infty), \mathbb{R}^{\mathfrak{d}})$ and

$$\dot{\Theta}_t = \mathcal{G}(\Theta_t). \quad (5.10)$$

Combining this with the hypothesis that $\mathcal{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$ and the chain rule ensures that for all $t \in (0, \infty)$ it holds that $\Theta \in C^2([0, \infty), \mathbb{R}^{\mathfrak{d}})$ and

$$\ddot{\Theta}_t = \mathcal{G}'(\Theta_t)\dot{\Theta}_t = \mathcal{G}'(\Theta_t)\mathcal{G}(\Theta_t). \quad (5.11)$$

Theorem 5.1.3 and (5.10) therefore imply that

$$\begin{aligned} \Theta_{T+\gamma} &= \Theta_T + \gamma \dot{\Theta}_T + \int_0^1 (1-r)\gamma^2 \ddot{\Theta}_{T+r\gamma} dr \\ &= \Theta_T + \gamma \mathcal{G}(\Theta_T) + \gamma^2 \int_0^1 (1-r) \mathcal{G}'(\Theta_{T+r\gamma}) \mathcal{G}(\Theta_{T+r\gamma}) dr. \end{aligned} \quad (5.12)$$

This and (5.7) demonstrate that

$$\begin{aligned}
 & \|\Theta_{T+\gamma} - \theta\|_2 \\
 &= \left\| \Theta_T + \gamma \mathcal{G}(\Theta_T) + \gamma^2 \int_0^1 (1-r) \mathcal{G}'(\Theta_{T+r\gamma}) \mathcal{G}(\Theta_{T+r\gamma}) dr - (\Theta_T + \gamma \mathcal{G}(\Theta_T)) \right\|_2 \\
 &\leq \gamma^2 \int_0^1 (1-r) \|\mathcal{G}'(\Theta_{T+r\gamma}) \mathcal{G}(\Theta_{T+r\gamma})\|_2 dr \\
 &\leq c^2 \gamma^2 \int_0^1 r dr = \frac{c^2 \gamma^2}{2} \leq c^2 \gamma^2.
 \end{aligned} \tag{5.13}$$

The proof of Lemma 5.1.4 is thus complete. \square

Corollary 5.1.5 (Local error of the Euler method for GF ODEs). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $x, y \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad \theta = \Theta_T - \gamma (\nabla \mathcal{L})(\Theta_T), \tag{5.14}$$

$$\|(\nabla \mathcal{L})(x)\|_2 \leq c, \quad \text{and} \quad \|(\text{Hess } \mathcal{L})(x)y\|_2 \leq c \|y\|_2 \tag{5.15}$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2 \gamma^2. \tag{5.16}$$

Proof of Corollary 5.1.5. Throughout this proof, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = -(\nabla \mathcal{L})(\theta). \tag{5.17}$$

Note that the fact that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds$, the fact that $\theta = \Theta_T + \gamma \mathcal{G}(\Theta_T)$, the fact that for all $x \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathcal{G}(x)\|_2 \leq c$, the fact that for all $x, y \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathcal{G}'(x)y\|_2 \leq c \|y\|_2$, and Lemma 5.1.4 imply that $\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2 \gamma^2$. The proof of Corollary 5.1.5 is thus complete. \square

5.1.3 Lyapunov-type stability for GD optimization

Corollary 4.4.3 in Section 4.4.2 and Corollary 4.4.4 in Chapter 4 above, in particular, illustrate how Lyapunov-type functions can be employed to establish convergence properties for GFs. Roughly speaking, the next two results, Proposition 5.1.6 and Corollary 5.1.7 below, are the time-discrete analogs of Corollary 4.4.3 and Corollary 4.4.4, respectively.

Proposition 5.1.6 (Lyapunov-type stability for discrete-time dynamical systems). *Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $c \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, c]$, let $V: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$, $\Phi: \mathbb{R}^{\mathfrak{d}} \times [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$, and $\varepsilon: [0, c] \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, c]$ that*

$$V(\Phi(\theta, t)) \leq \varepsilon(t)V(\theta), \quad (5.18)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Phi(\Theta_{n-1}, \gamma_n). \quad (5.19)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$V(\Theta_n) \leq \left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] V(\xi). \quad (5.20)$$

Proof of Proposition 5.1.6. We prove (5.20) by induction on $n \in \mathbb{N}_0$. For the base case $n = 0$ note that the assumption that $\Theta_0 = \xi$ ensures that $V(\Theta_0) = V(\xi)$. This establishes (5.20) in the base case $n = 0$. For the induction step observe that (5.19) and (5.18) ensure that for all $n \in \mathbb{N}_0$ with $V(\Theta_n) \leq \left(\prod_{k=1}^n \varepsilon(\gamma_k) \right) V(\xi)$ it holds that

$$\begin{aligned} V(\Theta_{n+1}) &= V(\Phi(\Theta_n, \gamma_{n+1})) \leq \varepsilon(\gamma_{n+1})V(\Theta_n) \\ &\leq \varepsilon(\gamma_{n+1}) \left(\left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] V(\xi) \right) = \left[\prod_{k=1}^{n+1} \varepsilon(\gamma_k) \right] V(\xi). \end{aligned} \quad (5.21)$$

Induction thus establishes (5.20). The proof of Proposition 5.1.6 is thus complete. \square

Corollary 5.1.7 (On quadratic Lyapunov-type functions for the GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta, \xi \in \mathbb{R}^{\mathfrak{d}}$, $c \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, c]$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\|\cdot\|: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ be a norm, let $\varepsilon: [0, c] \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, c]$ that*

$$\|\theta - t(\nabla \mathcal{L})(\theta) - \vartheta\|^2 \leq \varepsilon(t)\|\theta - \vartheta\|^2, \quad (5.22)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathcal{L})(\Theta_{n-1}). \quad (5.23)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\| \leq \left[\prod_{k=1}^n [\varepsilon(\gamma_k)]^{1/2} \right] \|\xi - \vartheta\|. \quad (5.24)$$

Proof of Corollary 5.1.7. Throughout this proof, let $V: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\Phi: \mathbb{R}^d \times [0, \infty) \rightarrow \mathbb{R}^d$ satisfy for all $\theta \in \mathbb{R}^d$, $t \in [0, \infty)$ that

$$V(\theta) = \|\theta - \vartheta\|^2 \quad \text{and} \quad \Phi(\theta, t) = \theta - t(\nabla \mathcal{L})(\theta). \quad (5.25)$$

Observe that Proposition 5.1.6 (applied with $V \curvearrowright V$, $\Phi \curvearrowright \Phi$ in the notation of Proposition 5.1.6) and (5.25) imply that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|^2 = V(\Theta_n) \leq \left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] V(\xi) = \left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] \|\xi - \vartheta\|^2. \quad (5.26)$$

This establishes (5.24). The proof of Corollary 5.1.7 is thus complete. \square

Corollary 5.1.7, in particular, illustrates that the one-step Lyapunov stability assumption in (5.22) may provide us suitable estimates for the approximation errors associated to the GD optimization method; see (5.24) above. The next result, Lemma 5.1.8 below, now provides us sufficient conditions which ensure that the one-step Lyapunov stability condition in (5.22) is satisfied so that we are in the position to apply Corollary 5.1.7 above to obtain estimates for the approximation errors associated to the GD optimization method. Lemma 5.1.8 employs the growth condition and the coercivity-type condition. Results similar to Lemma 5.1.8 can, for example, be found in [1, Remark 2.1] and [3, Lemma 2.1]. We will employ the statement of Lemma 5.1.8 in our error analysis for the GD optimization method in Section 5.1.4 below.

Lemma 5.1.8 (Sufficient conditions for a one-step Lyapunov-type stability condition).

Let $\mathfrak{d} \in \mathbb{N}$, let $\langle \cdot, \cdot \rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $v \in \mathbb{R}^d$ that $\|v\| = \sqrt{\langle v, v \rangle}$, and let $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^d$, $\mathbb{B} = \{w \in \mathbb{R}^d: \|w - \vartheta\| \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^d, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\| \leq L \|\theta - \vartheta\|. \quad (5.27)$$

Then

(i) it holds that $c \leq L$,

(ii) it holds for all $\theta \in \mathbb{B}$, $\gamma \in [0, \infty)$ that

$$\|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|^2 \leq (1 - 2\gamma c + \gamma^2 L^2) \|\theta - \vartheta\|^2, \quad (5.28)$$

(iii) it holds for all $\gamma \in (0, \frac{2c}{L^2})$ that $0 \leq 1 - 2\gamma c + \gamma^2 L^2 < 1$, and

(iv) it holds for all $\theta \in \mathbb{B}$, $\gamma \in [0, \frac{c}{L^2}]$ that

$$\|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|^2 \leq (1 - c\gamma) \|\theta - \vartheta\|^2. \quad (5.29)$$

Proof of Lemma 5.1.8. First of all, note that (5.27) ensures that for all $\theta \in \mathbb{B}$, $\gamma \in [0, \infty)$ it holds that

$$\begin{aligned}
 0 \leq \|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|^2 &= \|(\theta - \vartheta) - \gamma(\nabla \mathcal{L})(\theta)\|^2 \\
 &= \|\theta - \vartheta\|^2 - 2\gamma \langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle + \gamma^2 \|(\nabla \mathcal{L})(\theta)\|^2 \\
 &\leq \|\theta - \vartheta\|^2 - 2\gamma c \|\theta - \vartheta\|^2 + \gamma^2 L^2 \|\theta - \vartheta\|^2 \\
 &= (1 - 2\gamma c + \gamma^2 L^2) \|\theta - \vartheta\|^2.
 \end{aligned} \tag{5.30}$$

This establishes item (ii). Moreover, note that the fact that $\mathbb{B} \setminus \{\vartheta\} \neq \emptyset$ and (5.30) assure that for all $\gamma \in [0, \infty)$ it holds that

$$1 - 2\gamma c + \gamma^2 L^2 \geq 0. \tag{5.31}$$

Hence, we obtain that

$$\begin{aligned}
 1 - \frac{c^2}{L^2} &= 1 - \frac{2c^2}{L^2} + \frac{c^2}{L^2} = 1 - 2\left[\frac{c}{L^2}\right]c + \left[\frac{c^2}{L^4}\right]L^2 \\
 &= 1 - 2\left[\frac{c}{L^2}\right]c + \left[\frac{c}{L^2}\right]^2 L^2 \geq 0.
 \end{aligned} \tag{5.32}$$

This implies that $\frac{c^2}{L^2} \leq 1$. Therefore, we obtain that $c^2 \leq L^2$. This establishes item (i). Furthermore, observe that (5.31) ensures that for all $\gamma \in (0, \frac{2c}{L^2})$ it holds that

$$0 \leq 1 - 2\gamma c + \gamma^2 L^2 = 1 - \underbrace{\gamma}_{>0} \underbrace{(2c - \gamma L^2)}_{>0} < 1. \tag{5.33}$$

This proves item (iii). In addition, note that for all $\gamma \in [0, \frac{c}{L^2}]$ it holds that

$$1 - 2\gamma c + \gamma^2 L^2 \leq 1 - 2\gamma c + \gamma \left[\frac{c}{L^2}\right] L^2 = 1 - c\gamma. \tag{5.34}$$

Combining this with (5.30) establishes item (iv). The proof of Lemma 5.1.8 is thus complete. \square

5.1.4 Error analysis for GD optimization

In this subsection we provide an error analysis for the GD optimization method. In particular, we show under suitable hypotheses (cf. Proposition 5.1.9 below) that the considered GD process converges to a local minimum point of the objective function of the considered optimization problem.

5.1.4.1 Error estimates for GD optimization

Proposition 5.1.9 (Error estimates for the GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{2c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (5.35)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (5.36)$$

Then

- (i) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds for all $n \in \mathbb{N}$ that $0 \leq 1 - 2c\gamma_n + (\gamma_n)^2 L^2 \leq 1$,
- (iii) it holds for all $n \in \mathbb{N}$ that $\|\Theta_n - \vartheta\|_2 \leq (1 - 2c\gamma_n + (\gamma_n)^2 L^2)^{1/2} \|\Theta_{n-1} - \vartheta\|_2 \leq r$,
- (iv) it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^n (1 - 2c\gamma_k + (\gamma_k)^2 L^2)^{1/2} \right] \|\xi - \vartheta\|_2, \quad (5.37)$$

and

- (v) it holds for all $n \in \mathbb{N}_0$ that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{L}{2} \left[\prod_{k=1}^n (1 - 2c\gamma_k + (\gamma_k)^2 L^2) \right] \|\xi - \vartheta\|_2^2. \quad (5.38)$$

Proof of Proposition 5.1.9. First, note that (5.35) and item (ii) in Lemma 4.4.7 prove item (i). Moreover, observe that (5.35), item (iii) in Lemma 5.1.8, the assumption that for all $n \in \mathbb{N}$ it holds that $\gamma_n \in [0, \frac{2c}{L^2}]$, and the fact that

$$1 - 2c \left[\frac{2c}{L^2} \right] + \left[\frac{2c}{L^2} \right]^2 L^2 = 1 - \frac{4c^2}{L^2} + \left[\frac{4c^2}{L^4} \right] L^2 = 1 - \frac{4c^2}{L^2} + \frac{4c^2}{L^2} = 1 \quad (5.39)$$

and establish item (ii). Next we claim that for all $n \in \mathbb{N}$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq (1 - 2c\gamma_n + (\gamma_n)^2 L^2)^{1/2} \|\Theta_{n-1} - \vartheta\|_2 \leq r. \quad (5.40)$$

We now prove (5.40) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (5.36), the

assumption that $\Theta_0 = \xi \in \mathbb{B}$, item (ii) in Lemma 5.1.8, and item (ii) ensure that

$$\begin{aligned} \|\Theta_1 - \vartheta\|_2^2 &= \|\Theta_0 - \gamma_1(\nabla\mathcal{L})(\Theta_0) - \vartheta\|_2^2 \\ &\leq (1 - 2c\gamma_1 + (\gamma_1)^2L^2)\|\Theta_0 - \vartheta\|_2^2 \\ &\leq \|\Theta_0 - \vartheta\|_2^2 \leq r^2. \end{aligned} \quad (5.41)$$

This establishes (5.40) in the base case $n = 1$. For the induction step observe that (5.36), item (ii) in Lemma 5.1.8, and item (ii) imply that for all $n \in \mathbb{N}$ with $\Theta_n \in \mathbb{B}$ it holds that

$$\begin{aligned} \|\Theta_{n+1} - \vartheta\|_2^2 &= \|\Theta_n - \gamma_{n+1}(\nabla\mathcal{L})(\Theta_n) - \vartheta\|_2^2 \\ &\leq \underbrace{(1 - 2c\gamma_{n+1} + (\gamma_{n+1})^2L^2)}_{\in[0,1]}\|\Theta_n - \vartheta\|_2^2 \\ &\leq \|\Theta_n - \vartheta\|_2^2 \leq r^2. \end{aligned} \quad (5.42)$$

This demonstrates that for all $n \in \mathbb{N}$ with $\|\Theta_n - \vartheta\|_2 \leq r$ it holds that

$$\|\Theta_{n+1} - \vartheta\|_2 \leq (1 - 2c\gamma_{n+1} + (\gamma_{n+1})^2L^2)^{1/2}\|\Theta_n - \vartheta\|_2 \leq r. \quad (5.43)$$

Induction thus proves (5.40). Next note that (5.40) establishes item (iii). Moreover, observe that induction, item (ii), and item (iii) prove item (iv). Furthermore, note that item (iii) and the fact that $\Theta_0 = \xi \in \mathbb{B}$ ensure that for all $n \in \mathbb{N}_0$ it holds that $\Theta_n \in \mathbb{B}$. Combining this, (5.35), and Lemma 4.4.8 with items (i) and (iv) establishes item (v). The proof of Proposition 5.1.9 is thus complete. \square

5.1.4.2 Size of the learning rates

In the next result, Corollary 5.1.10 below, we, roughly speaking, specialize Proposition 5.1.9 to the case where the learning rates $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{2c}{L^2}]$ are a constant sequence.

Corollary 5.1.10 (Convergence of GD for constant learning rates). *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\gamma \in (0, \frac{2c}{L^2})$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla\mathcal{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla\mathcal{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2, \quad (5.44)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma(\nabla\mathcal{L})(\Theta_{n-1}). \quad (5.45)$$

Then

(i) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,

(ii) it holds that $0 \leq 1 - 2c\gamma + \gamma^2L^2 < 1$,

(iii) it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq [1 - 2c\gamma + \gamma^2 L^2]^{n/2} \|\xi - \vartheta\|_2, \quad (5.46)$$

and

(iv) it holds for all $n \in \mathbb{N}_0$ that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{L}{2} [1 - 2c\gamma + \gamma^2 L^2]^n \|\xi - \vartheta\|_2^2. \quad (5.47)$$

Proof of Corollary 5.1.10. Observe that item (iii) in Lemma 5.1.8 proves item (ii). In addition, note that Proposition 5.1.9 establishes items (i), (iii), and (iv). The proof of Corollary 5.1.10 is thus complete. \square

Corollary 5.1.10 above establishes under suitable hypotheses convergence of the considered GD process in the case where the learning rates are constant and strictly smaller than $\frac{2c}{L^2}$. The next result, Theorem 5.1.11 below, demonstrates that the condition that the learning rates are strictly smaller than $\frac{2c}{L^2}$ in Corollary 5.1.10 can, in general, not be relaxed.

Theorem 5.1.11 (Sharp bounds on the learning rate for the convergence of GD).

Let $\mathfrak{d} \in \mathbb{N}$, $\alpha \in (0, \infty)$, $\gamma \in \mathbb{R}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\xi \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{\alpha}{2} \|\theta - \vartheta\|_2^2, \quad (5.48)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma(\nabla \mathcal{L})(\Theta_{n-1}) \quad (5.49)$$

(cf. Definition 3.3.4). Then

(i) it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle = \alpha \|\theta - \vartheta\|_2^2$,

(ii) it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\|(\nabla \mathcal{L})(\theta)\|_2 = \alpha \|\theta - \vartheta\|_2$,

(iii) it holds for all $n \in \mathbb{N}_0$ that $\|\Theta_n - \vartheta\|_2 = |1 - \gamma\alpha|^n \|\xi - \vartheta\|_2$, and

(iv) it holds that

$$\liminf_{n \rightarrow \infty} \|\Theta_n - \vartheta\|_2 = \limsup_{n \rightarrow \infty} \|\Theta_n - \vartheta\|_2 = \begin{cases} 0 & : \gamma \in (0, 2/\alpha) \\ \|\xi - \vartheta\|_2 & : \gamma \in \{0, 2/\alpha\} \\ \infty & : \gamma \in \mathbb{R} \setminus [0, 2/\alpha] \end{cases} . \quad (5.50)$$

Proof of Theorem 5.1.11. First of all, observe that for all $\theta \in \mathbb{R}^d$ it holds that $\mathcal{L} \in C^\infty(\mathbb{R}^d, \mathbb{R})$ and

$$(\nabla \mathcal{L})(\theta) = \frac{\alpha}{2}(2(\theta - \vartheta)) = \alpha(\theta - \vartheta). \quad (5.51)$$

This proves item (ii). Moreover, observe that (5.51) assures that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle = \langle \theta - \vartheta, \alpha(\theta - \vartheta) \rangle = \alpha \|\theta - \vartheta\|_2^2. \quad (5.52)$$

This establishes item (i). Note that (5.49) and (5.51) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Theta_n - \vartheta &= \Theta_{n-1} - \gamma(\nabla \mathcal{L})(\Theta_{n-1}) - \vartheta \\ &= \Theta_{n-1} - \gamma\alpha(\Theta_{n-1} - \vartheta) - \vartheta \\ &= (1 - \gamma\alpha)(\Theta_{n-1} - \vartheta). \end{aligned} \quad (5.53)$$

The assumption that $\Theta_0 = \xi$ and induction hence prove that for all $n \in \mathbb{N}_0$ it holds that

$$\Theta_n - \vartheta = (1 - \gamma\alpha)^n(\Theta_0 - \vartheta) = (1 - \gamma\alpha)^n(\xi - \vartheta). \quad (5.54)$$

Therefore, we obtain for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 = |1 - \gamma\alpha|^n \|\xi - \vartheta\|_2. \quad (5.55)$$

This establishes item (iii). Combining item (iii) with the fact that for all $t \in (0, 2/\alpha)$ it holds that $|1 - t\alpha| \in [0, 1)$, the fact that for all $t \in \{0, 2/\alpha\}$ it holds that $|1 - t\alpha| = 1$, the fact that for all $t \in \mathbb{R} \setminus [0, 2/\alpha]$ it holds that $|1 - t\alpha| \in (1, \infty)$, and the fact that $\|\xi - \vartheta\|_2 > 0$ establishes item (iv). The proof of Theorem 5.1.11 is thus complete. \square

Exercise 5.1.3. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 2\theta^2 \quad (5.56)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0 = 1$ and

$$\Theta_n = \Theta_{n-1} - n^{-2}(\nabla \mathcal{L})(\Theta_{n-1}). \quad (5.57)$$

Prove or disprove the following statement: It holds that

$$\limsup_{n \rightarrow \infty} |\Theta_n| = 0. \quad (5.58)$$

Exercise 5.1.4. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 4\theta^2 \quad (5.59)$$

and for every $r \in (1, \infty)$ let $\Theta^{(r)}: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0^{(r)} = 1$ and

$$\Theta_n^{(r)} = \Theta_{n-1}^{(r)} - n^{-r}(\nabla \mathcal{L})(\Theta_{n-1}^{(r)}). \quad (5.60)$$

Prove or disprove the following statement: It holds for all $r \in (1, \infty)$ that

$$\liminf_{n \rightarrow \infty} |\Theta_n^{(r)}| > 0. \quad (5.61)$$

Exercise 5.1.5. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 5\theta^2 \quad (5.62)$$

and for every $r \in (1, \infty)$ let $\Theta^{(r)} = (\Theta_n^{(r)})_{n \in \mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0^{(r)} = 1$ and

$$\Theta_n^{(r)} = \Theta_{n-1}^{(r)} - n^{-r}(\nabla \mathcal{L})(\Theta_{n-1}^{(r)}). \quad (5.63)$$

Prove or disprove the following statement: It holds for all $r \in (1, \infty)$ that

$$\liminf_{n \rightarrow \infty} |\Theta_n^{(r)}| > 0. \quad (5.64)$$

5.1.4.3 Convergence rates

The next result, Corollary 5.1.12 below, establishes a convergence rate for the GD optimization method in the case of possibly non-constant learning rates. We prove Corollary 5.1.12 through an application of Proposition 5.1.9 above.

Corollary 5.1.12 (Qualitative convergence of GD). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$, $c, L \in (0, \infty)$, $\xi, \vartheta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2, \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (5.65)$$

$$\text{and} \quad 0 < \liminf_{n \rightarrow \infty} \gamma_n \leq \limsup_{n \rightarrow \infty} \gamma_n < \frac{2c}{L^2}, \quad (5.66)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (5.67)$$

Then

(i) it holds that $\{\theta \in \mathbb{R}^{\mathfrak{d}}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(w)\} = \{\vartheta\}$,

(ii) there exist $\epsilon \in (0, 1)$, $C \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \epsilon^n C, \quad (5.68)$$

and

(iii) there exist $\epsilon \in (0, 1)$, $C \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \epsilon^n C. \quad (5.69)$$

Proof of Corollary 5.1.12. Throughout this proof, let $\alpha, \beta \in \mathbb{R}$ satisfy

$$0 < \alpha < \liminf_{n \rightarrow \infty} \gamma_n \leq \limsup_{n \rightarrow \infty} \gamma_n < \beta < \frac{2c}{L^2} \quad (5.70)$$

(cf. (5.66)), let $m \in \mathbb{N}$ satisfy for all $n \in \mathbb{N}$ that $\gamma_{m+n} \in [\alpha, \beta]$, and let $h: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $t \in \mathbb{R}$ that

$$h(t) = 1 - 2ct + t^2 L^2. \quad (5.71)$$

Observe that (5.65) and item (ii) in Lemma 4.4.7 prove item (i). In addition, observe that the fact that for all $t \in \mathbb{R}$ it holds that $h'(t) = -2c + 2tL^2$ implies that for all $t \in (-\infty, \frac{c}{L^2}]$ it holds that

$$h'(t) \leq -2c + 2[\frac{c}{L^2}]L^2 = 0. \quad (5.72)$$

The fundamental theorem of calculus hence assures that for all $t \in [\alpha, \beta] \cap [0, \frac{c}{L^2}]$ it holds that

$$h(t) = h(\alpha) + \int_{\alpha}^t h'(s) ds \leq h(\alpha) + \int_{\alpha}^t 0 ds = h(\alpha) \leq \max\{h(\alpha), h(\beta)\}. \quad (5.73)$$

Furthermore, observe that the fact that for all $t \in \mathbb{R}$ it holds that $h'(t) = -2c + 2tL^2$ implies that for all $t \in [\frac{c}{L^2}, \infty)$ it holds that

$$h'(t) \leq h'(\frac{c}{L^2}) = -2c + 2[\frac{c}{L^2}]L^2 = 0. \quad (5.74)$$

The fundamental theorem of calculus hence ensures that for all $t \in [\alpha, \beta] \cap [\frac{c}{L^2}, \infty)$ it holds that

$$\max\{h(\alpha), h(\beta)\} \geq h(\beta) = h(t) + \int_t^{\beta} h'(s) ds \geq h(t) + \int_t^{\beta} 0 ds = h(t). \quad (5.75)$$

Combining this and (5.73) establishes that for all $t \in [\alpha, \beta]$ it holds that

$$h(t) \leq \max\{h(\alpha), h(\beta)\}. \quad (5.76)$$

Moreover, observe that the fact that $\alpha, \beta \in (0, \frac{2c}{L^2})$ and item (iii) in Lemma 5.1.8 ensure that

$$\{h(\alpha), h(\beta)\} \subseteq [0, 1). \quad (5.77)$$

Hence, we obtain that

$$\max\{h(\alpha), h(\beta)\} \in [0, 1). \quad (5.78)$$

This implies that there exists $\varepsilon \in \mathbb{R}$ such that

$$0 \leq \max\{h(\alpha), h(\beta)\} < \varepsilon < 1. \quad (5.79)$$

Next note that the fact that for all $n \in \mathbb{N}$ it holds that $\gamma_{m+n} \in [\alpha, \beta] \subseteq [0, \frac{2c}{L^2}]$, items (ii) and (iv) in Proposition 5.1.9 (applied with $\mathfrak{d} \curvearrowright \mathfrak{d}$, $c \curvearrowright c$, $L \curvearrowright L$, $r \curvearrowright \infty$, $(\gamma_n)_{n \in \mathbb{N}} \curvearrowright$

$(\gamma_{m+n})_{n \in \mathbb{N}}$, $\vartheta \curvearrowright \vartheta$, $\xi \curvearrowright \Theta_m$, $\mathcal{L} \curvearrowright \mathcal{L}$ in the notation of Proposition 5.1.9), (5.65), (5.67), and (5.76) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \|\Theta_{m+n} - \vartheta\|_2 &\leq \left[\prod_{k=1}^n (1 - 2c\gamma_{m+k} + (\gamma_{m+k})^2 L^2)^{1/2} \right] \|\Theta_m - \vartheta\|_2 \\ &= \left[\prod_{k=1}^n (h(\gamma_{m+k}))^{1/2} \right] \|\Theta_m - \vartheta\|_2 \\ &\leq (\max\{h(\alpha), h(\beta)\})^{n/2} \|\Theta_m - \vartheta\|_2 \\ &\leq \varepsilon^{n/2} \|\Theta_m - \vartheta\|_2. \end{aligned} \quad (5.80)$$

This shows that for all $n \in \mathbb{N}$ with $n > m$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \varepsilon^{(n-m)/2} \|\Theta_m - \vartheta\|_2. \quad (5.81)$$

The fact that for all $n \in \mathbb{N}_0$ with $n \leq m$ it holds that

$$\|\Theta_n - \vartheta\|_2 = \left[\frac{\|\Theta_n - \vartheta\|_2}{\varepsilon^{n/2}} \right] \varepsilon^{n/2} \leq \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\} \right] \varepsilon^{n/2} \quad (5.82)$$

hence assures that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\leq \max \left\{ \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\} \right] \varepsilon^{n/2}, \varepsilon^{(n-m)/2} \|\Theta_m - \vartheta\|_2 \right\} \\ &= (\varepsilon^{1/2})^n \left[\max \left\{ \max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\}, \varepsilon^{-m/2} \|\Theta_m - \vartheta\|_2 \right\} \right] \\ &= (\varepsilon^{1/2})^n \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\} \right]. \end{aligned} \quad (5.83)$$

This proves item (ii). In addition, note that Lemma 4.4.8, item (i), and (5.83) assure that for all $n \in \mathbb{N}_0$ it holds that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{\varepsilon^n L}{2} \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2^2}{\varepsilon^k} : k \in \{0, 1, \dots, m\} \right\} \right]. \quad (5.84)$$

This establishes item (iii). The proof of Corollary 5.1.12 is thus complete. \square

5.1.4.4 Error estimates in the case of small learning rates

The inequality in (5.37) in item (iv) in Proposition 5.1.9 above provides us an error estimate for the GD optimization method in the case where the learning rates $(\gamma_n)_{n \in \mathbb{N}}$ in

Proposition 5.1.9 satisfy that for all $n \in \mathbb{N}$ it holds that $\gamma_n \leq \frac{2c}{L^2}$. The error estimate in (5.37) can be simplified in the special case where the learning rates $(\gamma_n)_{n \in \mathbb{N}}$ satisfy the more restrictive condition that for all $n \in \mathbb{N}$ it holds that $\gamma_n \leq \frac{c}{L^2}$. This is the subject of the next result, Corollary 5.1.13 below. We prove Corollary 5.1.13 through an application of Proposition 5.1.9 above.

Corollary 5.1.13 (Error estimates in the case of small learning rates). *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (5.85)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (5.86)$$

Then

- (i) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds for all $n \in \mathbb{N}$ that $0 \leq 1 - c\gamma_n \leq 1$,
- (iii) it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^n (1 - c\gamma_k)^{1/2} \right] \|\xi - \vartheta\|_2 \leq \exp\left(-\frac{c}{2} \left[\sum_{k=1}^n \gamma_k \right]\right) \|\xi - \vartheta\|_2, \quad (5.87)$$

and

- (iv) it holds for all $n \in \mathbb{N}_0$ that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \left[\prod_{k=1}^n (1 - c\gamma_k) \right] \|\xi - \vartheta\|_2^2 \leq \frac{L}{2} \exp\left(-c \left[\sum_{k=1}^n \gamma_k \right]\right) \|\xi - \vartheta\|_2^2. \quad (5.88)$$

Proof of Corollary 5.1.13. Note that item (ii) in Proposition 5.1.9 and the assumption that for all $n \in \mathbb{N}$ it holds that $\gamma_n \in [0, \frac{c}{L^2}]$ ensure that for all $n \in \mathbb{N}$ it holds that

$$0 \leq 1 - 2c\gamma_n + (\gamma_n)^2 L^2 \leq 1 - 2c\gamma_n + \gamma_n \left[\frac{c}{L^2} \right] L^2 = 1 - 2c\gamma_n + \gamma_n c = 1 - c\gamma_n \leq 1. \quad (5.89)$$

This proves item (ii). Moreover, note that (5.89) and Proposition 5.1.9 establish items (i), (iii), and (iv). The proof of Corollary 5.1.13 is thus complete. \square

In the next result, Corollary 5.1.14 below, we, roughly speaking, specialize Corollary 5.1.13 above to the case where the learning rates $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{c}{L^2}]$ are a constant sequence.

Corollary 5.1.14 (Error estimates in the case of small and constant learning rates). *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\gamma \in (0, \frac{c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (5.90)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma (\nabla \mathcal{L})(\Theta_{n-1}). \quad (5.91)$$

Then

- (i) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds that $0 \leq 1 - c\gamma < 1$,
- (iii) it holds for all $n \in \mathbb{N}_0$ that $\|\Theta_n - \vartheta\|_2 \leq (1 - c\gamma)^{n/2} \|\xi - \vartheta\|_2$, and
- (iv) it holds for all $n \in \mathbb{N}_0$ that $0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} (1 - c\gamma)^n \|\xi - \vartheta\|_2^2$.

Proof of Corollary 5.1.14. Corollary 5.1.14 is an immediate consequence of Corollary 5.1.13. The proof of Corollary 5.1.14 is thus complete. \square

Chapter 6

Stochastic gradient descent (SGD) optimization methods

This chapter reviews and studies the classical plain-vanilla SGD optimization method (see Section 6.2). More sophisticated SGD-type optimization methods are SGD-type optimization methods with momenta and SGD-type optimization methods with adaptive modifications of the learning rates.

6.1 Introductory comments for the training of ANNs with SGD

In Chapter 5 we have introduced and studied deterministic GD-type optimization methods. In deep learning algorithms usually not deterministic GD-type optimization methods but stochastic variants of GD-type optimization methods are employed. Such SGD-type optimization methods can be viewed as suitable Monte Carlo approximations of deterministic GD-type methods and in this section we now roughly sketch some of the main ideas of such SGD-type optimization methods. To do this, we now briefly recall the deep supervised learning framework developed in the [introduction](#) and Section 4.1 above.

Specifically, let $d, M \in \mathbb{N}$, $\mathcal{E} \in C(\mathbb{R}^d, \mathbb{R})$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{M+1} \in \mathbb{R}^d$, $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}$ satisfy for all $m \in \{1, 2, \dots, M\}$ that

$$\mathbf{y}_m = \mathcal{E}(\mathbf{x}_m). \quad (6.1)$$

As in the [introduction](#) and in Section 4.1 we think of $M \in \mathbb{N}$ as the number of available known input-output data pairs, we think of $d \in \mathbb{N}$ as the dimension of the input data, we think of $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ as an unknown function which we want to approximate, we think of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{M+1} \in \mathbb{R}^d$ as the available known input data, we think of $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}$ as the available known output data, and we are trying to use the available known input-output data pairs to approximate the unknown function \mathcal{E} by means of ANNs.

Specifically, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \dots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1} + 1)] + l_h + 1$, and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \left| \mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d}(x_m) - y_m \right|^2 \right] \quad (6.2)$$

(cf. Definitions 1.1.3 and 1.2.1). Note that h is the number of hidden layers of the ANNs in (6.2), note for every $i \in \{1, 2, \dots, h\}$ that $l_i \in \mathbb{N}$ is the number of neurons in the i -th hidden layer of the ANNs in (6.2), and note that \mathfrak{d} is the number of real parameters used to describe the ANNs in (6.2). We recall that we are trying to approximate the function \mathcal{E} by, first, computing an approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ of the function $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ and, thereafter, employing the realization

$$\mathbb{R}^d \ni x \mapsto \mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\vartheta, d} \in \mathbb{R} \quad (6.3)$$

of the ANN associated to the approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ as an approximation of \mathcal{E} .

Deep learning algorithms typically solve optimization problems of the type (6.2) by means of gradient based optimization methods, which aim to minimize the considered objective function by performing successive steps based on the direction of the negative gradient of the objective function. We recall that one of the simplest gradient based optimization method is the plain-vanilla GD optimization method which performs successive steps in the direction of the negative gradient. In the context of the optimization problem in (6.2) this GD optimization method reads as follows. Let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, and let $\theta = (\theta_n)_{n \in \mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\theta_0 = \xi \quad \text{and} \quad \theta_n = \theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\theta_{n-1}). \quad (6.4)$$

Note that the process $(\theta_n)_{n \in \mathbb{N}_0}$ is the GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ (cf. Definition 5.1.1). Moreover, observe that the assumption that a is differentiable ensures that \mathcal{L} in (6.4) is also differentiable (see ?? above for details).

In typical practical deep learning applications the number M of available known input-output data pairs is very large, say, for instance, $M \geq 10^6$. As a consequence it is typically computationally prohibitively expensive to determine the exact gradient of the objective function to perform steps of deterministic GD-type optimization methods. As a remedy for this, deep learning algorithms usually employ stochastic variants of GD-type optimization methods, where in each step of the optimization method the precise gradient of the objective function is replaced by a Monte Carlo approximation of the gradient of the objective function. We now sketch this approach for the GD optimization method in (6.4) resulting in the popular SGD optimization method applied to (6.2).

Specifically, let $S = \{1, 2, \dots, M\}$, $J \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J\}$ let $\mathfrak{m}_{n,j}: \Omega \rightarrow S$ be a uniformly distributed random variable, let

$\ell: \mathbb{R}^d \times S \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$, $m \in S$ that

$$\ell(\theta, m) = \left| \left(\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d} \right) (x_m) - y_m \right|^2, \quad (6.5)$$

and let $\Theta = (\Theta_n)_{n \in \mathbb{N}_0}: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J} \sum_{j=1}^J (\nabla_{\theta} \ell)(\Theta_{n-1}, \mathfrak{m}_{n,j}) \right]. \quad (6.6)$$

The stochastic process $(\Theta_n)_{n \in \mathbb{N}_0}$ is an SGD process for the minimization problem associated to (6.2) with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, constant number of Monte Carlo samples (batch sizes) J , initial value ξ , and data $(\mathfrak{m}_{n,j})_{(n,j) \in \mathbb{N} \times \{1,2,\dots,J\}}$ (see Definition 6.2.1 below for the precise definition). Note that in (6.6) in each step $n \in \mathbb{N}$ we only employ a Monte Carlo approximation

$$\frac{1}{J} \sum_{j=1}^J (\nabla_{\theta} \ell)(\Theta_{n-1}, \mathfrak{m}_{n,j}) \approx \frac{1}{M} \sum_{m=1}^M (\nabla_{\theta} \ell)(\Theta_{n-1}, m) = (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.7)$$

of the exact gradient of the objective function. Nonetheless, in deep learning applications the SGD optimization method (or other SGD-type optimization methods) typically result in good approximate minimizers of the objective function. Note that employing approximate gradients in the SGD optimization method in (6.6) means that performing any step of the SGD process involves the computation of a sum with only J summands, while employing the exact gradient in the GD optimization method in (6.4) means that performing any step of the process involves the computation of a sum with M summands. In deep learning applications when M is very large (for example, $M \geq 10^6$) and J is chosen to be reasonably small (for instance, $J = 128$), this means that performing steps of the SGD process is much more computationally affordable than performing steps of the GD process. Combining this with the fact that SGD-type optimization methods do in the training of ANNs often find good approximate minimizers is the key reason making the SGD optimization method and other SGD-type optimization methods the optimization methods chosen in almost all deep learning applications. It is the topic of this chapter to introduce and study SGD-type optimization methods such as the plain-vanilla SGD optimization method in (6.6) above.

6.2 SGD optimization

In the next notion we present the promised stochastic version of the plain-vanilla GD optimization method from Section 5.1, that is, in the next notion we present the plain-vanilla SGD optimization method.

Definition 6.2.1 (SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let (S, \mathcal{S}) be a measurable space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J_n\}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ and $\mathfrak{g}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $U \in \{V \subseteq \mathbb{R}^{\mathfrak{d}}: V \text{ is open}\}$, $x \in S$, $\theta \in U$ with $(U \ni \vartheta \mapsto \ell(\vartheta, x) \in \mathbb{R}) \in C^1(U, \mathbb{R})$ that

$$\mathfrak{g}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x). \quad (6.8)$$

Then we say that Θ is the SGD process on $((\Omega, \mathcal{F}, \mathbb{P}), (S, \mathcal{S}))$ for the loss function ℓ with generalized gradient \mathfrak{g} , learning rates $(\gamma_n)_{n \in \mathbb{N}}$, batch sizes $(J_n)_{n \in \mathbb{N}}$, initial value ξ , and data $(X_{n,j})_{(n,j) \in \{(k,l) \in \mathbb{N}^2: l \leq J_k\}}$ if $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ is the function from $\mathbb{N}_0 \times \Omega$ to $\mathbb{R}^{\mathfrak{d}}$ which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathfrak{g}(\Theta_{n-1}, X_{n,j}) \right]. \quad (6.9)$$

6.2.1 SGD optimization in the training of ANNs

In the next example we apply the SGD optimization method in the context of the training of ANNs in the vectorized description (see Section 1.1) with the loss function being the mean squared error loss function in Definition 4.2.2 (see Section 4.2.2). Note that this is a very similar framework as the one developed in Section 6.1.

Example 6.2.2. Let $d, h, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_h \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1} + 1)] + l_h + 1$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M \in \mathbb{R}^d$, $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \left| (\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta, d}(\mathbf{x}_m) - \mathbf{y}_m \right|^2 \right], \quad (6.10)$$

let $S = \{1, 2, \dots, M\}$, let $\ell: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $m \in S$ that

$$\ell(\theta, m) = \left| (\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta, d}(\mathbf{x}_m) - \mathbf{y}_m \right|^2, \quad (6.11)$$

let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $\vartheta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\vartheta_0 = \xi \quad \text{and} \quad \vartheta_n = \vartheta_{n-1} - \gamma_n (\nabla \mathcal{L})(\vartheta_{n-1}), \quad (6.12)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, for every $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J_n\}$ let $\mathbf{m}_{n,j}: \Omega \rightarrow S$ be a uniformly distributed random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta_{n-1}, \mathbf{m}_{n,j}) \right]. \quad (6.13)$$

Then

- (i) it holds that ϑ is the GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ ,
- (ii) it holds that Θ is the SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, batch sizes $(J_n)_{n \in \mathbb{N}}$, initial value ξ , and data $(\mathbf{m}_{n,j})_{(n,j) \in \{(k,l) \in \mathbb{N}^2: l \leq J_k\}}$, and
- (iii) it holds for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that

$$\mathbb{E} \left[\theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\theta, \mathbf{m}_{n,j}) \right] \right] = \theta - \gamma_n (\nabla \mathcal{L})(\theta). \quad (6.14)$$

Proof for Example 6.2.2. Observe that (6.12) shows item (i). Note that (6.13) proves item (ii). Observe that (6.11), (6.10), and the assumption that for all $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J_n\}$ it holds that $\mathbf{m}_{n,j}$ is uniformly distributed prove that for all $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J_n\}$ it holds that

$$\begin{aligned} \mathbb{E}[\ell(\eta, \mathbf{m}_{n,j})] &= \frac{1}{M} \left[\sum_{m=1}^M \ell(\eta, m) \right] \\ &= \frac{1}{M} \left[\sum_{m=1}^M |(\mathcal{N}_{\mathfrak{m}_{a,l_1}, \mathfrak{m}_{a,l_2}, \dots, \mathfrak{m}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta, d}(x_m) - y_m|^2 \right] = \mathcal{L}(\theta). \end{aligned} \quad (6.15)$$

Hence, we obtain for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that

$$\begin{aligned} \mathbb{E} \left[\theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\theta, \mathbf{m}_{n,j}) \right] \right] &= \theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathbb{E}[(\nabla_{\theta} \ell)(\theta, \mathbf{m}_{n,j})] \right] \\ &= \theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla \mathcal{L})(\theta) \right] \\ &= \theta - \gamma_n (\nabla \mathcal{L})(\theta). \end{aligned} \quad (6.16)$$

The proof for Example 6.2.2 is thus complete. \square

Figures 6.1 and 6.2 show the approximations of the respective target functions by the realization functions of the ANNs at various points during the training.

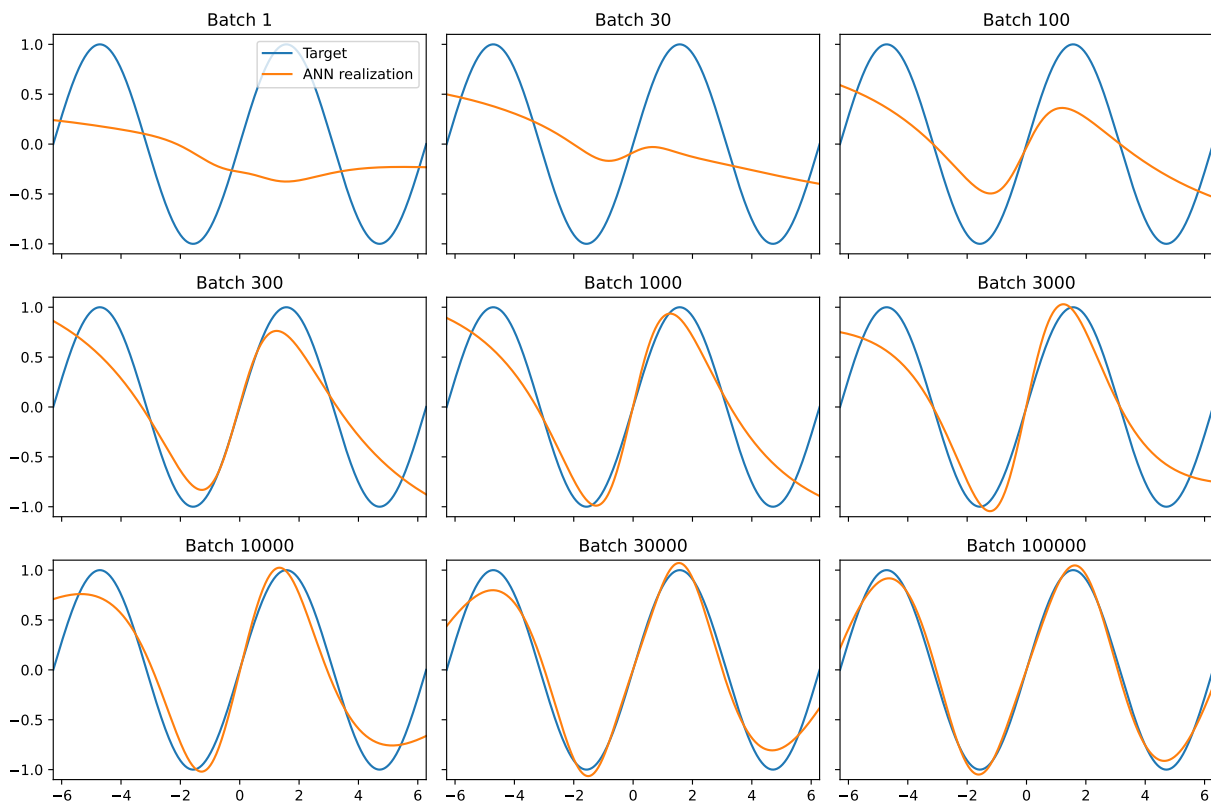


Figure 6.1 ([plots/sgd.pdf](#)): A plot showing the realization function of an ANN at several points during training with the SGD optimization method. The ANN with a single hidden layer with 200 neurons using the hyperbolic tangent activation function is trained so that the realization function approximates the target function $\sin: \mathbb{R} \rightarrow \mathbb{R}$. Example 6.2.2 is implemented with $d = 1$, $h = 1$, $\mathfrak{d} = 301$, $l_1 = 200$, $a = \tanh$, $M = 10000$, $x_1, x_2, \dots, x_M \in \mathbb{R}$, $y_i = \sin(x_i)$ for all $i \in \{1, 2, \dots, M\}$, $\gamma_n = 0.003$ for all $n \in \mathbb{N}$, and $J_n = 32$ for all $n \in \mathbb{N}$ in the notation of Example 6.2.2.

6.2.2 Convergence rates for SGD for coercive objective functions

The statement and the proof of the next result, Theorem 6.2.3 below, can be found in Jentzen et al. [3, Theorem 1.1].

Theorem 6.2.3. Let $\mathfrak{d} \in \mathbb{N}$, $p, \alpha, \kappa, c \in (0, \infty)$, $\nu \in (0, 1)$, $q = \min(\{2, 4, 6, \dots\} \cap [p, \infty))$, $\xi, \vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let (S, \mathcal{S}) be a measurable space, let $X_n: \Omega \rightarrow S$, $n \in \mathbb{N}$, be i.i.d. random variables, let $\ell = (\ell(\theta, x))_{\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable, assume for all $x \in S$ that $(\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \ell(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, assume for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbb{E}[|\ell(\theta, X_1)| + \|(\nabla_{\theta}\ell)(\theta, X_1)\|_2] < \infty, \quad (6.17)$$

$$\langle \theta - \vartheta, \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_1)] \rangle \geq c \max\{\|\theta - \vartheta\|_2^2, \|\mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_1)]\|_2^2\}, \quad (6.18)$$

$$\text{and} \quad \mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_1) - \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_1)]\|_2^q] \leq \kappa(1 + \|\theta\|_2^q), \quad (6.19)$$

let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathcal{L}(\theta) = \mathbb{E}[\ell(\theta, X_1)]$, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\alpha}{n^{\nu}}(\nabla_{\theta}\ell)(\Theta_{n-1}, X_n). \quad (6.20)$$

Then

(i) it holds that $\{\theta \in \mathbb{R}^{\mathfrak{d}}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(w)\} = \{\vartheta\}$ and

(ii) there exists $c \in \mathbb{R}$ such that for all $n \in \mathbb{N}$ it holds that

$$(\mathbb{E}[\|\Theta_n - \vartheta\|_2^p])^{1/p} \leq cn^{-\nu/2}. \quad (6.21)$$

Proof of Theorem 6.2.3. Note that Jentzen et al. [3, Theorem 1.1] establishes items (i) and (ii). The proof of Theorem 6.2.3 is thus complete. \square

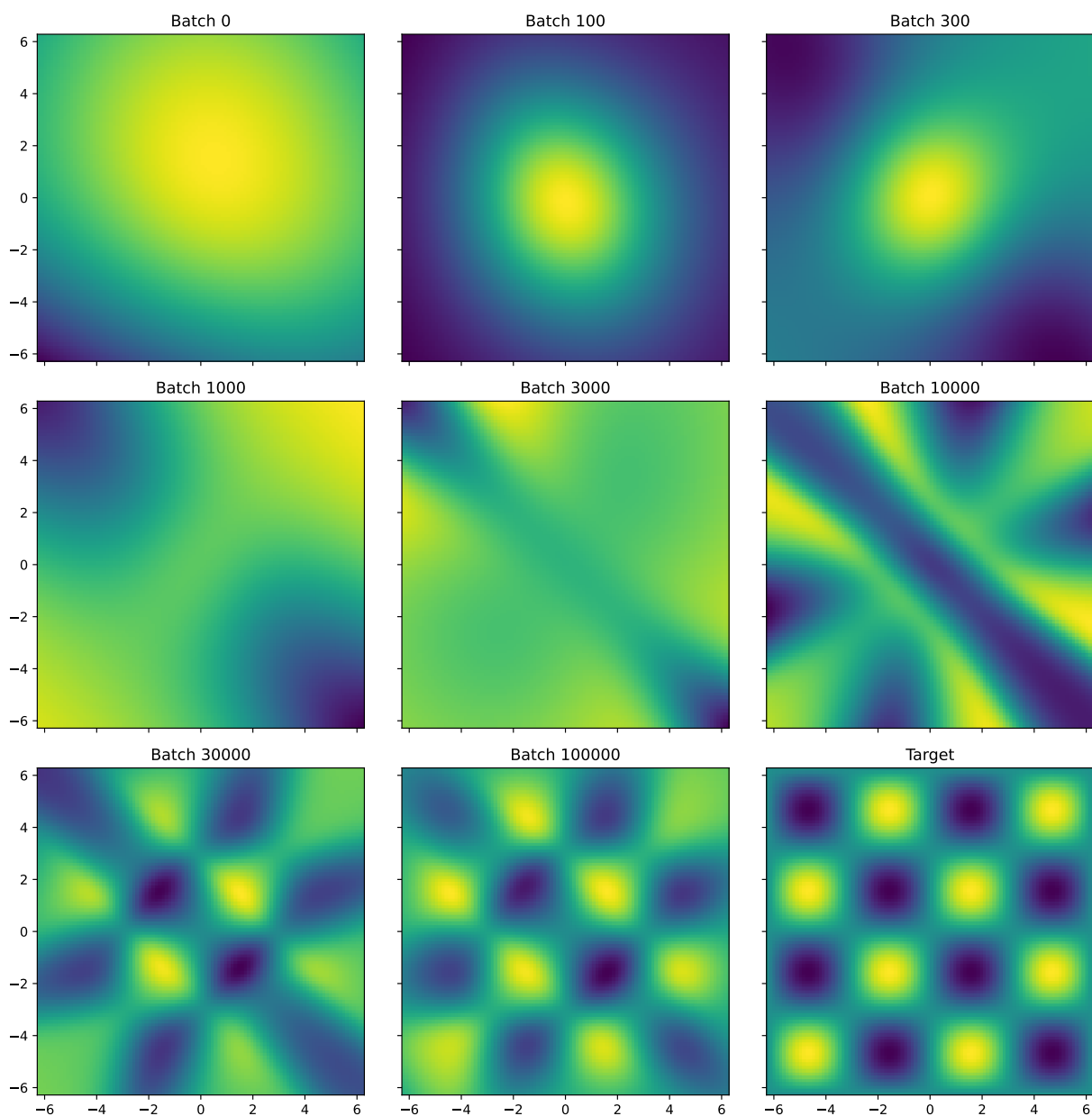


Figure 6.2 ([plots/sgd2.pdf](#)): A plot showing the realization function of an ANN at several points during training with the SGD optimization method. The ANN with two hidden layers with 50 neurons each using the softplus activation function is trained so that the realization function approximates the target function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ which satisfies for all $x, y \in \mathbb{R}$ that $f(x, y) = \sin(x) \sin(y)$. Example 6.2.2 is implemented with $d = 1$, $h = 2$, $\mathfrak{d} = 2701$, $l_1 = l_2 = 50$, a being the softplus activation function, $M = 10000$, $x_1, x_2, \dots, x_M \in \mathbb{R}^2$, $y_i = f(x_i)$ for all $i \in \{1, 2, \dots, M\}$, $\gamma_n = 0.003$ for all $n \in \mathbb{N}$, and $J_n = 32$ for all $n \in \mathbb{N}$ in the notation of Example 6.2.2.

Chapter 7

List of definitions

Chapter 1

Definition 1.1.1: Affine functions	14
Definition 1.1.3: Vectorized description of ANNs	15
Definition 1.2.1: Multidimensional versions of one-dimensional functions	19
Definition 1.2.4: ReLU activation function	21
Definition 1.2.5: Multidimensional ReLU activation functions	21
Definition 1.2.8: Softplus activation function	24
Definition 1.2.10: Hyperbolic tangent activation function	25
Definition 1.2.11: Softsign activation function	26
Definition 1.2.13: Leaky ReLU activation function	27
Definition 1.2.15: GELU activation function	28
Definition 1.2.17: Swish activation function	28
Definition 1.3.1: Structured description of ANNs	29
Definition 1.3.2: ANNs	30
Definition 1.3.4: Realizations of ANNs	31
Definition 1.3.6: Transformation from the structured to the vectorized description of ANNs	32

Chapter 2

Definition 2.1.1: Composition of ANNs	36
Definition 2.1.4: Powers of ANNs	39
Definition 2.2.1: Parallelization of ANNs	39
Definition 2.2.6: ReLU identity ANNs	44
Definition 2.2.8: Extensions of ANNs	46
Definition 2.2.12: Parallelization of ANNs with different length	49
Definition 2.3.1: Affine transformation ANNs	51
Definition 2.3.4: Scalar multiplications of ANNs	53
Definition 2.4.1: Sums of vectors as ANNs	54

Definition 2.4.5: Transpose of a matrix	56
Definition 2.4.6: Concatenation of vectors as ANNs.....	56
Definition 2.4.10: Sums of ANNs with the same length.....	58
Chapter 3	
Definition 3.1.1: Modulus of continuity	62
Definition 3.1.5: Linear interpolation operator	64
Definition 3.2.1: Identity matrices	68
Definition 3.2.2: Activation functions as ANNs.....	69
Definition 3.3.4: Quasi vector norms.....	78
Chapter 5	
Definition 5.1.1: GD optimization method	99
Chapter 6	
Definition 6.2.1: SGD optimization method.....	118

Bibliography

- [1] DEREICH, S. AND MÜLLER-GRONBACH, T. General multilevel adaptations for stochastic approximation algorithms. *arXiv:1506.05482* (2017), 33 pages. URL: arxiv.org/abs/1506.05482.
- [2] HENDRYCKS, D. AND GIMPEL, K. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415v4* (2016), 10 pp. URL: arxiv.org/abs/1606.08415.
- [3] JENTZEN, A., KUCKUCK, B., NEUFELD, A., AND VON WURSTEMBERGER, P. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA J. Numer. Anal.* 41, 1 (2020), pp. 455–492. URL: doi.org/10.1093/imanum/drz055.
- [4] JENTZEN, A., KUCKUCK, B., AND VON WURSTEMBERGER, P. Mathematical Introduction to Deep Learning: Methods, Implementations, and Theory. *arXiv:2310.20360* (2023), 601 pp. URL: arxiv.org/abs/2310.20360.
- [5] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *Nature* 521 (2015), pp. 436–444. URL: doi.org/10.1038/nature14539.