Mathematical Introduction to Deep Learning: Methods, Implementations, and Theory

Arnulf Jentzen
Benno Kuckuck
Philippe von Wurstemberger

Arnulf Jentzen

School of Data Science and Shenzhen Research Institute of Big Data
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)
Shenzhen, China
email: ajentzen@cuhk.edu.cn

Applied Mathematics: Institute for Analysis and Numerics
University of Münster
Münster, Germany
email: ajentzen@uni-muenster.de

Benno Kuckuck

School of Data Science and Shenzhen Research Institute of Big Data
The Chinese University of Hong Kong Shenzhen (CUHK-Shenzhen)
Shenzhen, China
email: bkuckuck@cuhk.edu.cn

Applied Mathematics: Institute for Analysis and Numerics
University of Münster
Münster, Germany
email: bkuckuck@uni-muenster.de

Philippe von Wurstemberger

School of Data Science
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)
Shenzhen, China
email: philippevw@cuhk.edu.cn

Risklab, Department of Mathematics
ETH Zurich
Zurich, Switzerland
email: philippe.vonwurstemberger@math.ethz.ch

Keywords: deep learning, artificial neural network, stochastic gradient descent, optimization
Mathematics Subject Classification (2020): 68T07

Version of Thursday 27th February, 2025

All PYTHON source codes in this book can be downloaded from

https://github.com/introdeeplearning/book

or from the arXiv page of this book (by clicking on "Other formats" and then "Download source").

Preface

This book aims to provide an introduction to the topic of deep learning algorithms. Very 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.

We hope that this book will be useful for students and scientists who do not yet have any background in deep learning at all and would like to gain a solid foundation as well as for practitioners who would like to obtain a firmer mathematical understanding of the objects and methods considered in deep learning.

After a brief introduction, this book is divided into six parts (see Parts I, II, III, IV, V, and VI). In Part I we introduce in Chapter 1 different types of ANNs including *fully-connected feedforward ANNs*, *convolutional ANNs* (CNNs), *recurrent ANNs* (RNNs), and *residual ANNs* (ResNets) in all mathematical details and in Chapter 2 we present a certain calculus for fully-connected feedforward ANNs.

In Part II we present several mathematical results that analyze how well ANNs can approximate given functions. To make this part more accessible, we first restrict ourselves in Chapter 3 to one-dimensional functions from the reals to the reals and, thereafter, we study ANN approximation results for multivariate functions in Chapter 4.

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 6 and stochastic variants of such gradient based optimization methods such as the *stochastic gradient descent* (SGD) optimization method are reviewed and studied in Chapter 7. 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 5 such GF ODEs. In particular, we show in Chapter 5 how such GF ODEs can be used to approximately solve appropriate optimization problems. Implementations of the gradient based methods discussed in Chapters 6 and 7 require efficient computations of gradients. The most popular and in some sense most natural method to explicitly compute such gradients in the case of the training of ANNs is the *backpropagation* method, which we derive and present in detail in Chapter 8. The mathematical analyses for gradient based optimization methods that we present in Chapters 5, 6, and 7 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 *Kurdyka–Łojasiewicz* (KL) approach which we discuss in detail in Chapter 9. In Chapter 10 we rigorously review *batch normalization* (BN) methods, which are popular methods that aim to accelerate ANN training procedures in data-driven learning problems. In Chapter 11 we review and study the approach to optimize an objective function through different random initializations.

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. It is precisely the subject of Part IV to study the generalization error. Specifically, in Chapter 12 we review suitable probabilistic generalization error estimates and in Chapter 13 we review suitable strong L^p-type generalization error estimates.

In Part V we illustrate how to combine parts of the *approximation error* estimates from Part II, parts of the *optimization error* estimates from Part III, and parts of the *generalization error* estimates from Part IV 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. Specifically, in Chapter 14 we present a suitable overall error decomposition for supervised learning problems, which we employ in Chapter 15 together with some of the findings of Parts II, III, and IV to establish the aforementioned illustrative overall error analysis.

Deep learning methods have not only become very popular for data-driven learning problems, but are nowadays also heavily used for approximately solving *partial differential equations* (PDEs). In Part VI we review and implement three popular variants of such deep learning methods for PDEs. Specifically, in Chapter 16 we treat *physics-informed neural networks* (PINNs) and *deep Galerkin methods* (DGMs) and in Chapter 17 we treat *deep Kolmogorov methods* (DKMs).

This book contains a number of PYTHON source codes, which can be downloaded from two sources, namely from the public GitHub repository at

https://github.com/introdeeplearning/book

and from the arXiv page of this book (by clicking on the link "Other formats" and then on

"Download source"). For ease of reference, the caption of each source listing in this book contains the filename of the corresponding source file.

This book grew out of a series of lectures held by the authors at ETH Zurich, University of Münster, and the Chinese University of Hong Kong, Shenzhen. It is in parts based on recent joint articles of Christian Beck, Sebastian Becker, Weinan E, Lukas Gonon, Robin Graeber, Philipp Grohs, Fabian Hornung, Martin Hutzenthaler, Nor Jaafari, Joshua Lee Padgett, Adrian Riekert, Diyora Salimova, Timo Welti, and Philipp Zimmermann with the authors of this book. We thank all of our aforementioned co-authors for very fruitful collaborations. Special thanks are due to Timo Welti for his permission to integrate slightly modified extracts of the article [240] into this book. We also thank Lukas Gonon, Timo Kröger, Siyu Liang, and Joshua Lee Padget for several insightful discussions and useful suggestions. Finally, we thank the students of the courses that we held on the basis of preliminary material of this book for bringing several typos to our notice.

This work has been partially funded by the National Science Foundation of China (NSFC) under grant number 12250610192. Moreover, this work was supported by the internal project fund from the Shenzhen Research Institute of Big Data under grant T00120220001. The first author gratefully acknowledges the support of the Cluster of Excellence EXC 2044-390685587, Mathematics Münster: Dynamics-Geometry-Structure funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Shenzhen and Münster,
Thursday 27$^{\text{th}}$ February, 2025

Arnulf Jentzen
Benno Kuckuck
Philippe von Wurstemberger

Contents

Introduction

Very roughly speaking, the field *deep learning* can be divided into three subfields, deep *supervised learning*, deep *unsupervised learning*, and deep *reinforcement learning*. 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

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

In the framework described in the previous sentence 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 $\mathcal{E} \colon \mathbb{R}^d \to \mathbb{R}$ 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 $\mathcal{E} \colon \mathbb{R}^d \to \mathbb{R}$ but instead by using the knowledge of the M input-output data pairs

$$(x_1, y_1) = (x_1, \mathcal{E}(x_1)),\ (x_2, y_2) = (x_2, \mathcal{E}(x_2)),\ \dots\ , (x_M, y_M) = (x_M, \mathcal{E}(x_M)) \in \mathbb{R}^d \times \mathbb{R}. \tag{2}$$

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

$$\mathfrak{L}(\phi) = \frac{1}{M} \left[\sum_{m=1}^{M} |\phi(x_m) - y_m|^2 \right]. \tag{3}$$

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

$$\mathfrak{L} \colon C(\mathbb{R}^d, \mathbb{R}) \to [0, \infty). \tag{4}$$

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 $\mathfrak{d} \in \mathbb{N}$, let $\psi = (\psi_\theta)_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to C(\mathbb{R}^d, \mathbb{R})$ be a function, and let $\mathscr{L} \colon \mathbb{R}^\mathfrak{d} \to [0, \infty)$ satisfy

$$\mathscr{L} = \mathfrak{L} \circ \psi. \tag{5}$$

We think of the set

$$\{\psi_\theta \colon \theta \in \mathbb{R}^\mathfrak{d}\} \subseteq C(\mathbb{R}^d, \mathbb{R}) \tag{6}$$

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

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

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 $\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$, $x \in \mathbb{R}$ it holds that

$$\psi_\theta(x) = \sum_{k=0}^{\mathfrak{d}-1} \theta_{k+1} x^k \tag{8}$$

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 $\{\psi_\theta \colon \theta \in \mathbb{R}^\mathfrak{d}\}$, that is, we consider the optimization problem of computing approximate minimizers of the function

$$\{\psi_\theta \colon \theta \in \mathbb{R}^\mathfrak{d}\} \ni \phi \mapsto \mathfrak{L}(\phi) = \frac{1}{M} \left[\sum_{m=1}^{M} |\phi(x_m) - y_m|^2 \right] \in [0, \infty). \tag{9}$$

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

$$\mathbb{R}^\mathfrak{d} \ni \theta \mapsto \mathscr{L}(\theta) = \mathfrak{L}(\psi_\theta) = \frac{1}{M} \left[\sum_{m=1}^{M} |\psi_\theta(x_m) - y_m|^2 \right] \in [0, \infty) \tag{10}$$

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 7 in Part III we present the most common variants of such SGD-type optimization algorithms. If $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ is an approximate minimizer of (10) in the sense that $\mathscr{L}(\vartheta) \approx \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathscr{L}(\theta)$, one then considers $\psi_\vartheta(x_{M+1})$ as an approximation

$$\psi_\vartheta(x_{M+1}) \approx \mathcal{E}(x_{M+1}) \tag{11}$$

of the unknown output $\mathcal{E}(x_{M+1})$ of the $(M+1)$-th input data x_{M+1}. We note that in deep supervised learning algorithms one typically aims to compute an approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ of (10) in the sense that $\mathscr{L}(\vartheta) \approx \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathscr{L}(\theta)$, which is, however, typically not a minimizer of (10) in the sense that $\mathscr{L}(\vartheta) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathscr{L}(\theta)$ (cf. Section 9.15).

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 5 of Part III for a survey of commonly used loss function in deep learning algorithms (see Section 5.4.2). We also refer to Chapter 9 for convergence results in the above framework where the parametrization function in (7) corresponds to *fully-connected feedforward* ANNs (see Section 9.15).

Part I

Artificial neural networks (ANNs)

Chapter 1

Basics on ANNs

In this chapter we review different types of architectures of ANNs such as fully-connected feedforward ANNs (see Sections 1.1 and 1.3), CNNs (see Section 1.4), ResNets (see Section 1.5), and RNNs (see Section 1.6), we review different types of popular activation functions used in applications such as the *rectified linear unit* (ReLU) activation (see Section 1.2.3), the *Gaussian error linear unit* (GELU) activation (see Section 1.2.6), and the standard logistic activation (see Section 1.2.7) among others, and we review 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).

In the literature different types of ANN architectures and activation functions have been reviewed in several excellent works; cf., for example, [4, 9, 39, 41, 61, 64, 99, 170, 188, 197, 381, 387, 403, 445] and the references therein. The specific presentation of Sections 1.1 and 1.3 is based on [19, 20, 25, 165, 186].

1.1 Fully-connected feedforward ANNs (vectorized description)

We start the mathematical content of this book with a review of fully-connected feedforward ANNs, the most basic type of ANNs. 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. 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 1.1.5 below).

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

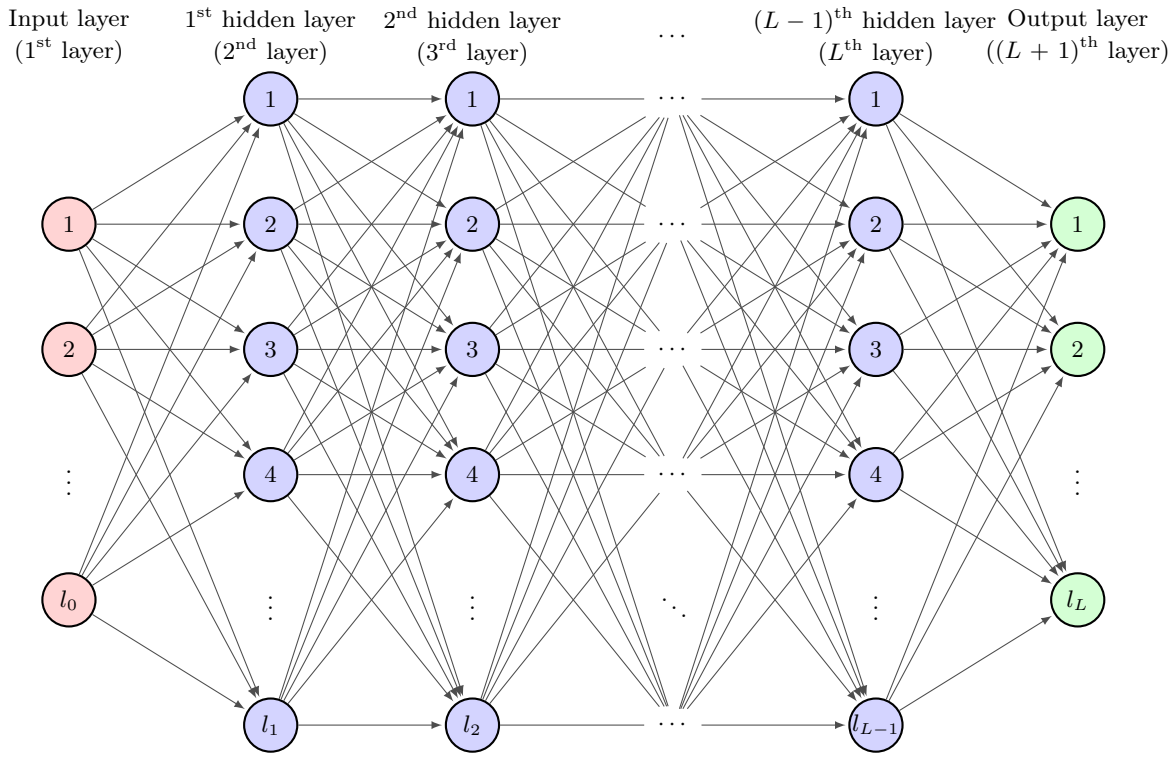

Figure 1.1: Graphical illustration of a fully-connected feedforward ANN consisting of $L \in \mathbb{N}$ affine transformations (i.e., consisting of $L+1$ layers: one input layer, $L-1$ hidden layers, and one output layer) with $l_0 \in \mathbb{N}$ neurons on the input layer (i.e., with l_0-dimensional input layer), with $l_1 \in \mathbb{N}$ neurons on the 1^{st} hidden layer (i.e., with l_1-dimensional 1^{st} hidden layer), with $l_2 \in \mathbb{N}$ neurons on the 2^{nd} hidden layer (i.e., with l_2-dimensional 2^{nd} hidden layer), ..., with l_{L-1} neurons on the $(L-1)^{\text{th}}$ hidden layer (i.e., with (l_{L-1})-dimensional $(L-1)^{\text{th}}$ 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 $\mathfrak{d}, m, n \in \mathbb{N}$, $s \in \mathbb{N}_0$, $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathfrak{d} \geq s + mn + m$. Then we denote by $\mathcal{A}^{\theta,s}_{m,n} \colon \mathbb{R}^n \to \mathbb{R}^m$ the function which satisfies for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ that*

$$\begin{aligned} \mathcal{A}^{\theta,s}_{m,n}(x) &= \begin{pmatrix} \theta_{s+1} & \theta_{s+2} & \cdots & \theta_{s+n} \\ \theta_{s+n+1} & \theta_{s+n+2} & \cdots & \theta_{s+2n} \\ \theta_{s+2n+1} & \theta_{s+2n+2} & \cdots & \theta_{s+3n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{s+(m-1)n+1} & \theta_{s+(m-1)n+2} & \cdots & \theta_{s+mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \theta_{s+mn+1} \\ \theta_{s+mn+2} \\ \theta_{s+mn+3} \\ \vdots \\ \theta_{s+mn+m} \end{pmatrix} \\ &= \Big(\big[\textstyle\sum_{k=1}^{n} x_k \theta_{s+k}\big] + \theta_{s+mn+1}, \big[\textstyle\sum_{k=1}^{n} x_k \theta_{s+n+k}\big] + \theta_{s+mn+2}, \ldots, \\ &\qquad \big[\textstyle\sum_{k=1}^{n} x_k \theta_{s+(m-1)n+k}\big] + \theta_{s+mn+m} \Big) \end{aligned} \tag{1.1}$$

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

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

(cf. Definition 1.1.1).

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

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

The proof for Example 1.1.2 is thus complete. □

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

1.1.2 Vectorized description of fully-connected feedforward ANNs

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

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

and for every $k \in \{1, 2, \ldots, L\}$ let $\Psi_k \colon \mathbb{R}^{l_k} \to \mathbb{R}^{l_k}$ be a function. Then we denote by

$\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L}\colon \mathbb{R}^{l_0} \to \mathbb{R}^{l_L}$ *the function given by*

$$\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L} = \Psi_L \circ \mathcal{A}^{\theta,\sum_{k=1}^{L-1} l_k(l_{k-1}+1)}_{l_L,l_{L-1}} \circ \Psi_{L-1} \circ \mathcal{A}^{\theta,\sum_{k=1}^{L-2} l_k(l_{k-1}+1)}_{l_{L-1},l_{L-2}} \circ \ldots \\ \ldots \circ \Psi_2 \circ \mathcal{A}^{\theta,l_1(l_0+1)}_{l_2,l_1} \circ \Psi_1 \circ \mathcal{A}^{\theta,0}_{l_1,l_0} \tag{1.5}$$

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

Example 1.1.4 (Example for Definition 1.1.3)**.** *Let* $\theta = (1, -1, 2, -2, 3, -3, 0, 0, 1) \in \mathbb{R}^9$ *and let* $\Psi\colon \mathbb{R}^2 \to \mathbb{R}^2$ *satisfy for all* $x = (x_1, x_2) \in \mathbb{R}^2$ *that*

$$\Psi(x) = (\max\{x_1, 0\}, \max\{x_2, 0\}). \tag{1.6}$$

Then

$$\left(\mathcal{N}^{\theta,1}_{\Psi,\mathrm{id}_{\mathbb{R}}}\right)(2) = 12 \tag{1.7}$$

(cf. Definition 1.1.3).

Proof for Example 1.1.4. Note that (1.1), (1.5), and (1.6) show that

$$\begin{aligned}\left(\mathcal{N}^{\theta,1}_{\Psi,\mathrm{id}_{\mathbb{R}}}\right)(2) &= \left(\mathrm{id}_{\mathbb{R}} \circ \mathcal{A}^{\theta,4}_{1,2} \circ \Psi \circ \mathcal{A}^{\theta,0}_{2,1}\right)(2) = \left(\mathcal{A}^{\theta,4}_{1,2} \circ \Psi\right)\left(\begin{pmatrix}1\\-1\end{pmatrix}(2) + \begin{pmatrix}2\\-2\end{pmatrix}\right) \\ &= \left(\mathcal{A}^{\theta,4}_{1,2} \circ \Psi\right)\left(\begin{pmatrix}4\\-4\end{pmatrix}\right) = \mathcal{A}^{\theta,4}_{1,2}\left(\begin{pmatrix}4\\0\end{pmatrix}\right) = \begin{pmatrix}3 & -3\end{pmatrix}\begin{pmatrix}4\\0\end{pmatrix} + (0) = 12\end{aligned} \tag{1.8}$$

(cf. Definitions 1.1.1 and 1.1.3). The proof for Example 1.1.4 is thus complete. □

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

$$\Psi(x) = (\max\{x_1, 0\}, \min\{x_2, 0\}). \tag{1.9}$$

Prove or disprove the following statement: It holds that

$$\left(\mathcal{N}^{\theta,1}_{\Psi,\mathrm{id}_{\mathbb{R}}}\right)(-1) = -1 \tag{1.10}$$

(cf. Definition 1.1.3).

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

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

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

$$m(x) = \max\{-x, 0\} \qquad \text{and} \qquad q(x) = x^2. \tag{1.11}$$

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

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

$$(\theta_1, \ldots, \theta_{15}) = (1, -2, 0, 3, 2, -1, 0, 3, 1, -1, 1, -1, 2, 0, -1) \tag{1.12}$$

and let $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ and $\Psi\colon \mathbb{R}^2 \to \mathbb{R}^2$ 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 $\left(\mathcal{N}^{\theta,2}_{\Phi,\Psi}\right)(1, -1) = (4, 4)$ (cf. Definition 1.1.3).

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

1.1.3 Weight and bias parameters of fully-connected feedforward ANNs

Remark 1.1.5 (Weights and biases for fully-connected feedforward ANNs). Let $L \in \{2, 3, 4, \ldots\}$, $v_0, v_1, \ldots, v_{L-1} \in \mathbb{N}_0$, $l_0, l_1, \ldots, l_L$, $\mathfrak{d} \in \mathbb{N}$, $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $k \in \{0, 1, \ldots, L-1\}$ that

$$\mathfrak{d} \geq \sum_{i=1}^{L} l_i(l_{i-1} + 1) \qquad \text{and} \qquad v_k = \sum_{i=1}^{k} l_i(l_{i-1} + 1), \tag{1.13}$$

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

$$W_k = \underbrace{\begin{pmatrix} \theta_{v_{k-1}+1} & \theta_{v_{k-1}+2} & \cdots & \theta_{v_{k-1}+l_{k-1}} \\ \theta_{v_{k-1}+l_{k-1}+1} & \theta_{v_{k-1}+l_{k-1}+2} & \cdots & \theta_{v_{k-1}+2l_{k-1}} \\ \theta_{v_{k-1}+2l_{k-1}+1} & \theta_{v_{k-1}+2l_{k-1}+2} & \cdots & \theta_{v_{k-1}+3l_{k-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{v_{k-1}+(l_k-1)l_{k-1}+1} & \theta_{v_{k-1}+(l_k-1)l_{k-1}+2} & \cdots & \theta_{v_{k-1}+l_k l_{k-1}} \end{pmatrix}}_{\text{weight parameters}} \tag{1.14}$$

$$\text{and} \qquad b_k = \underbrace{\left(\theta_{v_{k-1}+l_k l_{k-1}+1}, \theta_{v_{k-1}+l_k l_{k-1}+2}, \ldots, \theta_{v_{k-1}+l_k l_{k-1}+l_k}\right)}_{\text{bias parameters}}, \tag{1.15}$$

and let $\Psi_k\colon \mathbb{R}^{l_k} \to \mathbb{R}^{l_k}$, $k \in \{1, 2, \ldots, L\}$, be functions. Then

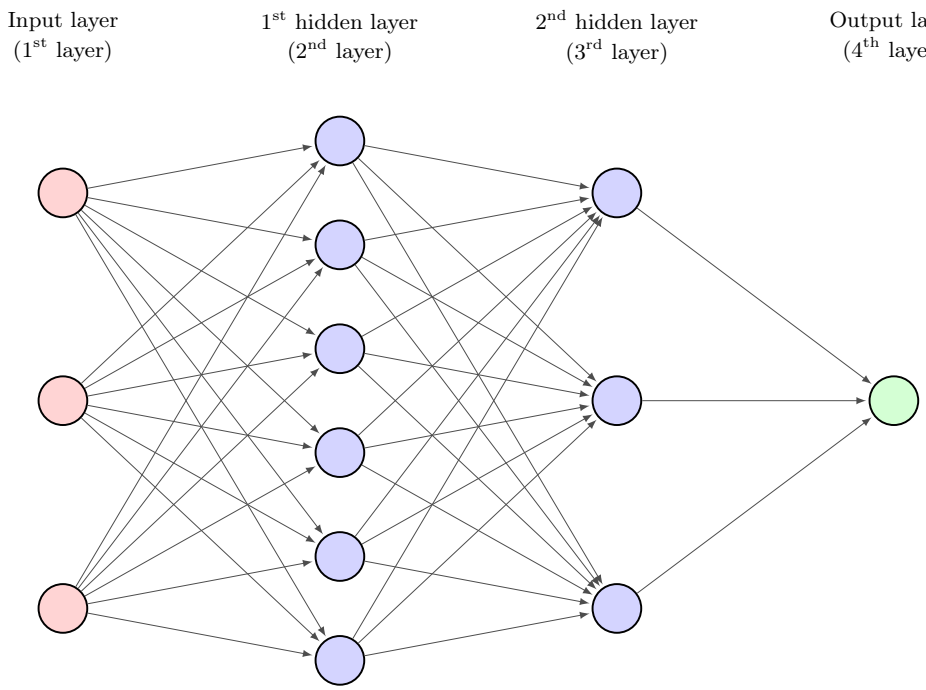

Figure 1.2: Graphical illustration of an ANN. The ANN has 2 hidden layers and length $L = 3$ with 3 neurons in the input layer (corresponding to $l_0 = 3$), 6 neurons in the first hidden layer (corresponding to $l_1 = 6$), 3 neurons in the second hidden layer (corresponding to $l_2 = 3$), and one neuron 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 $\mathbb{R}$.

(i) it holds that

$$\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L} = \Psi_L \circ \mathcal{A}^{\theta,v_{L-1}}_{l_L,l_{L-1}} \circ \Psi_{L-1} \circ \mathcal{A}^{\theta,v_{L-2}}_{l_{L-1},l_{L-2}} \circ \Psi_{L-2} \circ \ldots \circ \mathcal{A}^{\theta,v_1}_{l_2,l_1} \circ \Psi_1 \circ \mathcal{A}^{\theta,v_0}_{l_1,l_0} \tag{1.16}$$

and

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

(cf. Definitions 1.1.1 and 1.1.3).

1.2 Activation functions

In this section we review a few popular activation functions from the literature (cf. Definition 1.1.3 above and Definition 1.3.4 below for the use of activation functions in the context

of fully-connected feedforward ANNs, cf. Definition 1.4.5 below for the use of activation functions in the context of CNNs, cf. Definition 1.5.4 below for the use of activation functions in the context of ResNets, and cf. Definitions 1.6.3 and 1.6.4 below for the use of activation functions in the context of RNNs).

1.2.1 Multi-dimensional versions

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

Definition 1.2.1 (Multi-dimensional versions of one-dimensional functions)**.** *Let $T \in \mathbb{N}$, $d_1, d_2, \ldots, d_T \in \mathbb{N}$ and let $\psi \colon \mathbb{R} \to \mathbb{R}$ be a function. Then we denote by*

$$\mathfrak{M}_{\psi,d_1,d_2,\ldots,d_T} \colon \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_T} \to \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_T} \tag{1.17}$$

the function which satisfies for all $x = (x_{k_1,k_2,\ldots,k_T})_{(k_1,k_2,\ldots,k_T) \in (\times_{t=1}^{T} \{1,2,\ldots,d_t\})} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_T}$, $y = (y_{k_1,k_2,\ldots,k_T})_{(k_1,k_2,\ldots,k_T) \in (\times_{t=1}^{T} \{1,2,\ldots,d_t\})} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_T}$ with $\forall\, k_1 \in \{1,2,\ldots,d_1\}$, $k_2 \in \{1,2,\ldots,d_2\}$, $\ldots$, $k_T \in \{1,2,\ldots,d_T\} \colon y_{k_1,k_2,\ldots,k_T} = \psi(x_{k_1,k_2,\ldots,k_T})$ that

$$\mathfrak{M}_{\psi,d_1,d_2,\ldots,d_T}(x) = y \tag{1.18}$$

and we call $\mathfrak{M}_{\psi,d_1,d_2,\ldots,d_T}$ the $d_1 \times d_2 \times \ldots \times d_T$-dimensional version of ψ.

Example 1.2.2 (Example for Definition 1.2.1)**.** *Let $A \in \mathbb{R}^{3 \times 1 \times 2}$ satisfy*

$$A = \big(\begin{pmatrix} 1 & -1 \end{pmatrix}, \begin{pmatrix} -2 & 2 \end{pmatrix}, \begin{pmatrix} 3 & -3 \end{pmatrix} \big) \tag{1.19}$$

and let $\psi \colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that $\psi(x) = x^2$. Then

$$\mathfrak{M}_{\psi,3,1,2}(A) = \big(\begin{pmatrix} 1 & 1 \end{pmatrix}, \begin{pmatrix} 4 & 4 \end{pmatrix}, \begin{pmatrix} 9 & 9 \end{pmatrix} \big) \tag{1.20}$$

Proof for Example 1.2.2. Note that (1.18) establishes (1.20). The proof for Example 1.2.2 is thus complete. □

Exercise 1.2.1. Let $A \in \mathbb{R}^{2 \times 3}$, $B \in \mathbb{R}^{2 \times 2 \times 2}$ satisfy

$$A = \begin{pmatrix} 3 & -2 & 5 \\ 1 & 0 & -2 \end{pmatrix} \qquad \text{and} \qquad B = \left(\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} -3 & -4 \\ 5 & 2 \end{pmatrix} \right) \tag{1.21}$$

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

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

$$(\theta_1, \theta_2, \ldots, \theta_{14}) = (0, 1, 2, 2, 1, 0, 1, 1, 1, -3, -1, 4, 0, 1) \tag{1.22}$$

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

$$f(x) = \frac{1}{1 + |x|} \qquad \text{and} \qquad g(x) = x^2. \tag{1.23}$$

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

1.2.2 Single hidden layer fully-connected feedforward ANNs

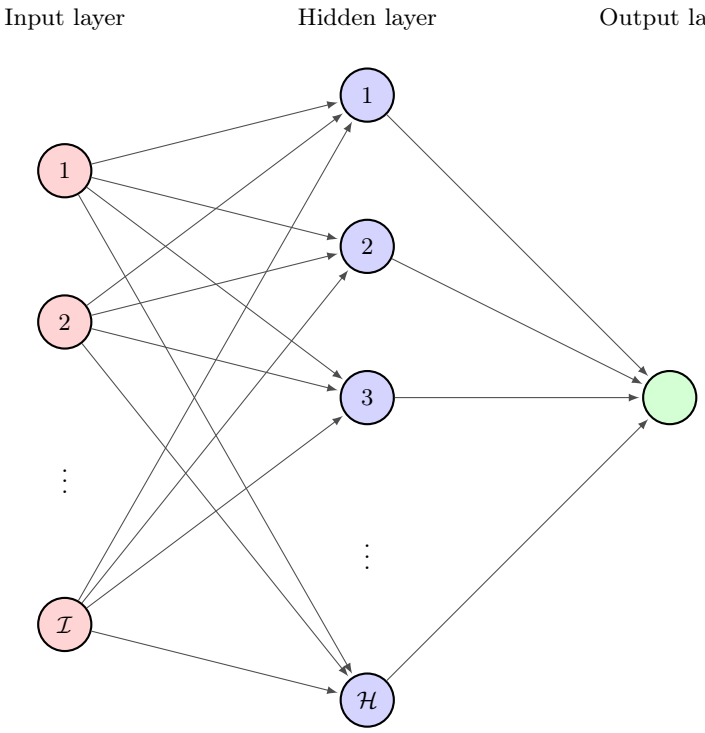

Figure 1.3: Graphical illustration of a fully-connected feedforward ANN consisting of two affine transformations (i.e., consisting of 3 layers: one input layer, one hidden layer, and one output layer) with $\mathcal{I} \in \mathbb{N}$ neurons on the input layer (i.e., with $\mathcal{I}$-dimensional input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

Lemma 1.2.3 (Fully-connected feedforward ANN with one hidden layer)**.** *Let $\mathcal{I}, \mathcal{H} \in \mathbb{N}$, $\theta = (\theta_1, \ldots, \theta_{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}) \in \mathbb{R}^{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}$, $x = (x_1, \ldots, x_{\mathcal{I}}) \in \mathbb{R}^{\mathcal{I}}$ and let $\psi \colon \mathbb{R} \to \mathbb{R}$ be a function. Then*

$$\mathcal{N}^{\theta,\mathcal{I}}_{\mathfrak{M}_{\psi,\mathcal{H}},\mathrm{id}_{\mathbb{R}}}(x) = \left[\sum_{k=1}^{\mathcal{H}} \theta_{\mathcal{H}\mathcal{I}+\mathcal{H}+k}\, \psi\left(\left[\sum_{i=1}^{\mathcal{I}} x_i \theta_{(k-1)\mathcal{I}+i}\right] + \theta_{\mathcal{H}\mathcal{I}+k}\right)\right] + \theta_{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}. \tag{1.24}$$

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

Proof of Lemma 1.2.3. Observe that (1.5) and (1.18) show that

$$\begin{aligned}
&\mathcal{N}^{\theta,\mathcal{I}}_{\mathfrak{M}_{\psi,\mathcal{H}},\mathrm{id}_{\mathbb{R}}}(x) \\
&= \left(\mathrm{id}_{\mathbb{R}} \circ \mathcal{A}^{\theta,\mathcal{H}\mathcal{I}+\mathcal{H}}_{1,\mathcal{H}} \circ \mathfrak{M}_{\psi,\mathcal{H}} \circ \mathcal{A}^{\theta,0}_{\mathcal{H},\mathcal{I}}\right)(x) \\
&= \mathcal{A}^{\theta,\mathcal{H}\mathcal{I}+\mathcal{H}}_{1,\mathcal{H}}\left(\mathfrak{M}_{\psi,\mathcal{H}}\left(\mathcal{A}^{\theta,0}_{\mathcal{H},\mathcal{I}}(x)\right)\right) \\
&= \left[\sum_{k=1}^{\mathcal{H}} \theta_{\mathcal{H}\mathcal{I}+\mathcal{H}+k}\, \psi\left(\left[\sum_{i=1}^{\mathcal{I}} x_i \theta_{(k-1)\mathcal{I}+i}\right] + \theta_{\mathcal{H}\mathcal{I}+k}\right)\right] + \theta_{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}.
\end{aligned} \tag{1.25}$$

The proof of Lemma 1.2.3 is thus complete. □

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. [273]).

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

$$\mathfrak{r}(x) = \max\{x, 0\} \tag{1.26}$$

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

```python
import matplotlib.pyplot as plt

def setup_axis(xlim, ylim):
    _, ax = plt.subplots()

    ax.set_aspect("equal")
    ax.set_xlim(xlim)
    ax.set_ylim(ylim)
    ax.spines["left"].set_position("zero")
    ax.spines["bottom"].set_position("zero")
```

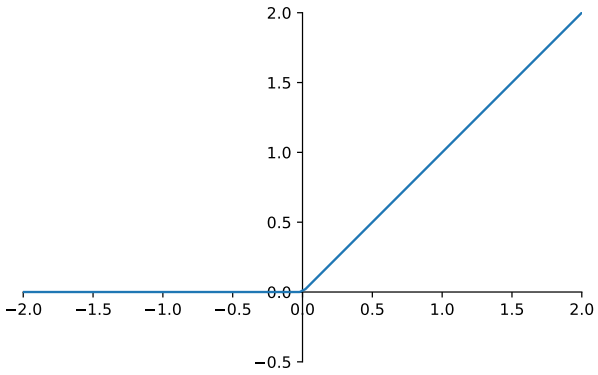

Figure 1.4 (plots/relu.pdf): A plot of the ReLU activation function

```
    ax.spines["right"].set_color("none")
    ax.spines["top"].set_color("none")
    for s in ax.spines.values():
        s.set_zorder(0)

    return ax
```

Source code 1.1 (code/activation_functions/plot_util.py): PYTHON code for the PLOT_UTIL module used in the code listings throughout this subsection

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-2,2), (-.5,2))

x = np.linspace(-2, 2, 100)

ax.plot(x, tf.keras.activations.relu(x))

plt.savefig("../../plots/relu.pdf", bbox_inches='tight')
```

Source code 1.2 (code/activation_functions/relu_plot.py): PYTHON code used to create Figure 1.4

Definition 1.2.5 (Multi-dimensional ReLU activation functions)**.** *Let $d \in \mathbb{N}$. Then we denote by $\mathfrak{R}_d \colon \mathbb{R}^d \to \mathbb{R}^d$ the function given by*

$$\mathfrak{R}_d = \mathfrak{M}_{\mathfrak{r},d} \tag{1.27}$$

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*

$$x = W_1 \max\{w_1 x + b_1, 0\} + W_2 \max\{w_2 x + b_2, 0\} + B. \tag{1.28}$$

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

$$\begin{aligned} &W_1 \max\{w_1 x + b_1, 0\} + W_2 \max\{w_2 x + b_2, 0\} + B \\ &= \max\{w_1 x + b_1, 0\} - \max\{w_2 x + b_2, 0\} = \max\{x, 0\} - \max\{-x, 0\} \\ &= \max\{x, 0\} + \min\{x, 0\} = x. \end{aligned} \tag{1.29}$$

The proof of Lemma 1.2.6 is thus complete. □

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

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \ldots, \mathfrak{R}_{l_H}, \mathrm{id}_{\mathbb{R}}}\right)(x) = x \tag{1.30}$$

(cf. Definitions 1.1.3 and 1.2.5).

The statement of the next lemma, Lemma 1.2.7, provides a partial answer to Exercise 1.2.3. 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.

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*

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_2, \mathrm{id}_{\mathbb{R}}}\right)(x) = x \tag{1.31}$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \ldots, \mathfrak{R}_{l_H}, \mathrm{id}_{\mathbb{R}}}\right)(x) = |x| \tag{1.32}$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \ldots, \mathfrak{R}_{l_H}, \mathrm{id}_{\mathbb{R}}}\right)(x) = e^x \tag{1.33}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.6 (Two-dimensional maximum). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 3l_1 + \big[\sum_{k=2}^{H} l_k(l_{k-1}+1)\big] + l_H + 1$ such that for all $x, y \in \mathbb{R}$ it holds that

$$\big(\mathcal{N}^{\theta,2}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_H},\mathrm{id}_{\mathbb{R}}}\big)(x,y) = \max\{x,y\} \tag{1.34}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.7 (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 $\mathfrak{d} \geq 2l_1 + l_1 l_2 + 2l_2 + 1$ such that for all $x \in \mathbb{R}$ it holds that

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

(cf. Definitions 1.1.3 and 1.2.5).

The statement of the next lemma, Lemma 1.2.8, provides a partial answer to Exercise 1.2.7. The proof of Lemma 1.2.8 is left as an exercise.

Lemma 1.2.8 (Real identity with two hidden layers). *Let* $\theta = (1, -1, 0, 0, 1, -1, -1, 1, 0, 0, 1, -1, 0) \in \mathbb{R}^{13}$. *Then it holds for all* $x \in \mathbb{R}$ *that*

$$\big(\mathcal{N}^{\theta,1}_{\mathfrak{R}_2,\mathfrak{R}_2,\mathrm{id}_{\mathbb{R}}}\big)(x) = x \tag{1.36}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.8 (Three-dimensional maximum). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 4l_1 + \big[\sum_{k=2}^{H} l_k(l_{k-1}+1)\big] + l_H + 1$ such that for all $x, y, z \in \mathbb{R}$ it holds that

$$\big(\mathcal{N}^{\theta,3}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_H},\mathrm{id}_{\mathbb{R}}}\big)(x,y,z) = \max\{x,y,z\} \tag{1.37}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.9 (Multi-dimensional maxima). Prove or disprove the following statement: For every $k \in \mathbb{N}$ there exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq (k+1)l_1 + \big[\sum_{k=2}^{H} l_k(l_{k-1}+1)\big] + l_H + 1$ such that for all $x_1, x_2, \ldots, x_k \in \mathbb{R}$ it holds that

$$\big(\mathcal{N}^{\theta,k}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_H},\mathrm{id}_{\mathbb{R}}}\big)(x_1,x_2,\ldots,x_k) = \max\{x_1,x_2,\ldots,x_k\} \tag{1.38}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.10. Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2\,l_1 + \big[\sum_{k=2}^{H} l_k(l_{k-1}+1)\big] + (l_H + 1)$ such that for all $x \in \mathbb{R}$ it holds that

$$\big(\mathcal{N}^{\theta,1}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_H},\mathrm{id}_{\mathbb{R}}}\big)(x) = \max\{x, \tfrac{x}{2}\} \tag{1.39}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.11 (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

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_l,\mathrm{id}_{\mathbb{R}}}\right)(x) = \begin{cases} 1 & : x \leq 2 \\ x - 1 & : 2 < x \leq 3 \\ 5 - x & : 3 < x \leq 4 \\ 1 & : x > 4 \end{cases} \tag{1.40}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.12. 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

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_l,\mathrm{id}_{\mathbb{R}}}\right)(x) = \begin{cases} -2 & : x \leq 1 \\ 2x - 4 & : 1 < x \leq 3 \\ 2 & : x > 3 \end{cases} \tag{1.41}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.13. Prove or disprove the following statement: There exists $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2\,l_1 + \left[\sum_{k=2}^{H} l_k(l_{k-1} + 1)\right] + (l_H + 1)$ such that for all $x \in \mathbb{R}$ it holds that

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_H},\mathrm{id}_{\mathbb{R}}}\right)(x) = \begin{cases} 0 & : x \leq 1 \\ x - 1 & : 1 \leq x \leq 2 \\ 1 & : x \geq 2 \end{cases} \tag{1.42}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.14. 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

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_l,\mathrm{id}_{\mathbb{R}}}\right)(x) = x^2 \tag{1.43}$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.15. Prove or disprove the following statement: There exists $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2\,l_1 + \left[\sum_{k=2}^{H} l_k(l_{k-1} + 1)\right] + (l_H + 1)$ such that

$$\sup_{x \in [-3,-2]} \left| \left(\mathcal{N}^{\theta,1}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_H},\mathrm{id}_{\mathbb{R}}}\right)(x) - (x+2)^2 \right| \leq \tfrac{1}{4} \tag{1.44}$$

(cf. Definitions 1.1.3 and 1.2.5).

1.2.4 Clipping activation

Definition 1.2.9 (Clipping activation functions)**.** *Let $u \in [-\infty, \infty)$, $v \in (u, \infty]$. Then we denote by $\mathfrak{c}_{u,v} \colon \mathbb{R} \to \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that*

$$\mathfrak{c}_{u,v}(x) = \max\{u, \min\{x, v\}\}. \tag{1.45}$$

and we call $\mathfrak{c}_{u,v}$ the (u, v)-clipping activation function.

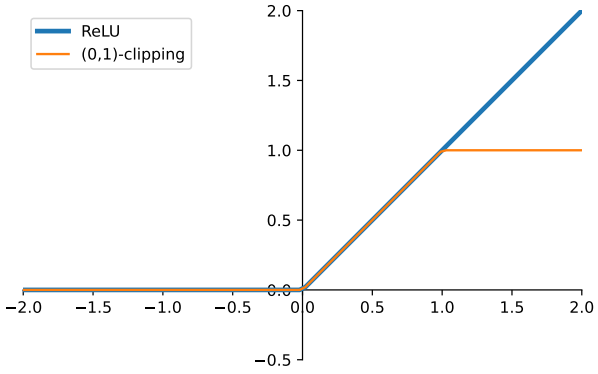

Figure 1.5 (plots/clipping.pdf): A plot of the $(0, 1)$-clipping activation function and the ReLU activation function

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-2,2), (-.5,2))

x = np.linspace(-2, 2, 100)

ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
ax.plot(x, tf.keras.activations.relu(x, max_value=1),
        label='(0,1)-clipping')
ax.legend()

plt.savefig("../../plots/clipping.pdf", bbox_inches='tight')
```

Source code 1.3 (code/activation_functions/clipping_plot.py): PYTHON code used to create Figure 1.5

Definition 1.2.10 (Multi-dimensional clipping activation functions)**.** *Let $d \in \mathbb{N}$, $u \in$*

$[-\infty, \infty)$, $v \in (u, \infty]$. *Then we denote by* $\mathfrak{C}_{u,v,d} \colon \mathbb{R}^d \to \mathbb{R}^d$ *the function given by*

$$\mathfrak{C}_{u,v,d} = \mathfrak{M}_{\mathfrak{c}_{u,v},d} \tag{1.46}$$

and we call $\mathfrak{C}_{u,v,d}$ *the* d*-dimensional* (u, v)*-clipping activation function (cf. Definitions 1.2.1 and 1.2.9).*

1.2.5 Softplus activation

Definition 1.2.11 (Softplus activation function)**.** *We say that* a *is the softplus activation function if and only if it holds that* $a \colon \mathbb{R} \to \mathbb{R}$ *is the function from* $\mathbb{R}$ *to* $\mathbb{R}$ *which satisfies for all* $x \in \mathbb{R}$ *that*

$$a(x) = \ln(1 + \exp(x)). \tag{1.47}$$

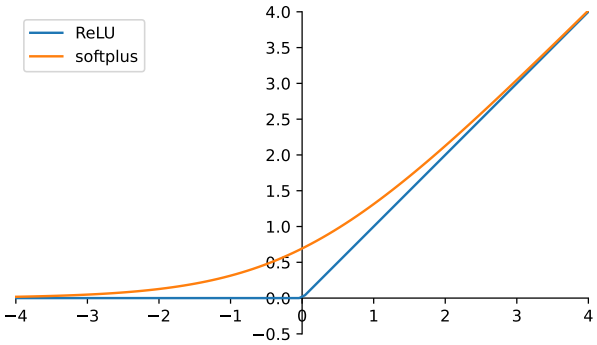

Figure 1.6 (plots/softplus.pdf): A plot of the softplus activation function and the ReLU activation function

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-4,4), (-.5,4))

x = np.linspace(-4, 4, 100)

ax.plot(x, tf.keras.activations.relu(x), label='ReLU')
ax.plot(x, tf.keras.activations.softplus(x), label='softplus')
ax.legend()

plt.savefig("../../plots/softplus.pdf", bbox_inches='tight')
```

Source code 1.4 (code/activation_functions/softplus_plot.py): PYTHON code used to create Figure 1.6

The next result, Lemma 1.2.12 below, presents a few elementary properties of the softplus function.

Lemma 1.2.12 (Properties of the softplus function)**.** *Let a be the softplus activation function (cf. Definition 1.2.11). Then*

(i) it holds for all $x \in [0, \infty)$ that $x \leq a(x) \leq x + 1$,

(ii) it holds that $\limsup_{x \to -\infty} a(x) = 0$,

(iii) it holds that $\limsup_{x \to \infty} a(x) = \infty$, and

(iv) it holds that $a(0) = \ln(2)$

(cf. Definition 1.2.11).

Proof of Lemma 1.2.12. Observe that the fact that $2 \leq \exp(1)$ ensures that for all $x \in [0, \infty)$ it holds that

$$\begin{aligned} x &= \ln(\exp(x)) \leq \ln(1 + \exp(x)) = \ln(\exp(0) + \exp(x)) \\ &\leq \ln(\exp(x) + \exp(x)) = \ln(2\exp(x)) \leq \ln(\exp(1)\exp(x)) \\ &= \ln(\exp(x + 1)) = x + 1. \end{aligned} \tag{1.48}$$

The proof of Lemma 1.2.12 is thus complete. □

Note that Lemma 1.2.12 ensures that $\mathfrak{s}(0) = \ln(2) = 0.693\ldots$ (cf. Definition 1.2.11). In the next step we introduce the multi-dimensional version of the softplus function (cf. Definitions 1.2.1 and 1.2.11 above).

Definition 1.2.13 (Multi-dimensional softplus activation functions)**.** *Let $d \in \mathbb{N}$ and let a be the softplus activation function (cf. Definition 1.2.11). Then we say that A is the d-dimensional softplus activation function if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

Lemma 1.2.14. *Let $d \in \mathbb{N}$ and let $A \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function. Then A is the d-dimensional softplus activation function if and only if it holds for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ that*

$$A(x) = (\ln(1 + \exp(x_1)), \ln(1 + \exp(x_2)), \ldots, \ln(1 + \exp(x_d))) \tag{1.49}$$

(cf. Definition 1.2.13).

Proof of Lemma 1.2.14. Throughout this proof, let a be the softplus activation function (cf. Definition 1.2.11). Note that (1.18) and (1.47) establish that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ it holds that

$$\mathfrak{M}_{a,d}(x) = (\ln(1 + \exp(x_1)), \ln(1 + \exp(x_2)), \ldots, \ln(1 + \exp(x_d))) \tag{1.50}$$

(cf. Definition 1.2.1). The fact that A is the d-dimensional softplus activation function (cf. Definition 1.2.13) if and only if $A = \mathfrak{M}_{a,d}$ hence implies (1.49). The proof of Lemma 1.2.14 is thus complete. □

Exercise 1.2.16 (Real identity). For every $d \in \mathbb{N}$ let A_d be the d-dimensional softplus activation function (cf. Definition 1.2.13). Prove or disprove the following statement: There exist $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \ldots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + \left[\sum_{k=2}^{H} l_k(l_{k-1} + 1)\right] + l_H + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$\left(\mathcal{N}^{\theta,1}_{A_{l_1}, A_{l_2}, \ldots, A_{l_H}, \mathrm{id}_{\mathbb{R}}}\right)(x) = x \tag{1.51}$$

(cf. Definition 1.1.3).

1.2.6 Gaussian error linear unit (GELU) activation

Another popular activation function is the GELU activation function first introduced in Hendrycks & Gimpel [201]. This activation function is the subject of the next definition.

Definition 1.2.15 (GELU activation function)**.** *We say that a is the GELU unit activation function (we say that a is the GELU activation function) if and only if it holds that $a \colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \frac{x}{\sqrt{2\pi}} \left[\int_{-\infty}^{x} \exp(-\tfrac{z^2}{2}) \, \mathrm{d}z \right]. \tag{1.52}$$

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-4,3), (-.5,3))

x = np.linspace(-4, 3, 100)

ax.plot(x, tf.keras.activations.relu(x), label='ReLU')
ax.plot(x, tf.keras.activations.softplus(x), label='softplus')
```

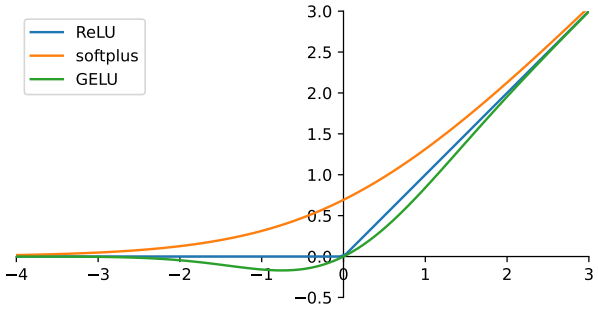

Figure 1.7 (plots/gelu.pdf): A plot of the GELU activation function, the ReLU activation function, and the softplus activation function

```
12 ax.plot(x, tf.keras.activations.gelu(x), label='GELU')
13 ax.legend()
14
15 plt.savefig("../../plots/gelu.pdf", bbox_inches='tight')
```

Source code 1.5 (code/activation_functions/gelu_plot.py): PYTHON code used to create Figure 1.7

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).

Proof of Lemma 1.2.16. Note that (1.26) and (1.52) imply that ((i) $\leftrightarrow$ (ii)). The proof of Lemma 1.2.16 is thus complete. □

Definition 1.2.17 (Multi-dimensional GELU activation functions)**.** *Let $d \in \mathbb{N}$ and let a be the GELU activation function (cf. Definition 1.2.15). Then we say that A is the d-dimensional GELU activation function if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.7 Standard logistic activation

Definition 1.2.18 (Standard logistic activation function)**.** *We say that a is the standard logistic activation function if and only if it holds that $a\colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$*

to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{\exp(x)+1}. \tag{1.53}$$

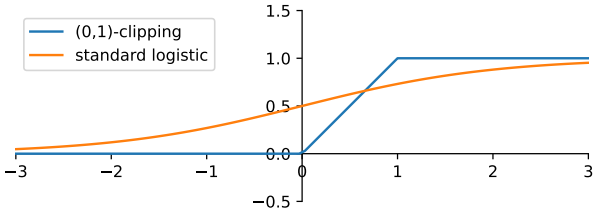

Figure 1.8 (plots/logistic.pdf): A plot of the standard logistic activation function and the $(0, 1)$-clipping activation function

```python
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-3,3), (-.5,1.5))

x = np.linspace(-3, 3, 100)

ax.plot(x, tf.keras.activations.relu(x, max_value=1),
        label='(0,1)-clipping')
ax.plot(x, tf.keras.activations.sigmoid(x),
        label='standard logistic')
ax.legend()

plt.savefig("../../plots/logistic.pdf", bbox_inches='tight')
```

Source code 1.6 (code/activation_functions/logistic_plot.py): PYTHON code used to create Figure 1.8

Definition 1.2.19 (Multi-dimensional standard logistic activation functions)**.** *Let $d \in \mathbb{N}$ and let a be the standard logistic activation function (cf. Definition 1.2.18). Then we say that A is the d-dimensional standard logistic activation function if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.7.1 Derivative of the standard logistic activation function

Proposition 1.2.20 (Logistic ODE). *Let a be the standard logistic activation function (cf. Definition 1.2.18). Then*

(i) it holds that $a\colon \mathbb{R} \to \mathbb{R}$ is infinitely often differentiable and

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

$$a(0) = 1/2, \qquad a'(x) = a(x)(1 - a(x)) = a(x) - [a(x)]^2, \qquad \text{and} \tag{1.54}$$

$$a''(x) = a(x)(1 - a(x))(1 - 2\,a(x)) = 2[a(x)]^3 - 3[a(x)]^2 + a(x). \tag{1.55}$$

Proof of Proposition 1.2.20. Note that (1.53) implies item (i). Next observe that (1.53) ensures that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} a'(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2} = a(x)\left(\frac{\exp(-x)}{1+\exp(-x)}\right) \\ &= a(x)\left(\frac{1+\exp(-x)-1}{1+\exp(-x)}\right) = a(x)\left(1 - \frac{1}{1+\exp(-x)}\right) \\ &= a(x)(1-a(x)). \end{aligned} \tag{1.56}$$

Hence, we obtain that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} a''(x) &= \big[a(x)(1-a(x))\big]' = a'(x)(1-a(x)) + a(x)(1-a(x))' \\ &= a'(x)(1-a(x)) - a(x)\,a'(x) = a'(x)(1-2\,a(x)) \\ &= a(x)(1-a(x))(1-2\,a(x)) \\ &= \big(a(x) - [a(x)]^2\big)(1-2\,a(x)) = a(x) - [a(x)]^2 - 2[a(x)]^2 + 2[a(x)]^3 \\ &= 2[a(x)]^3 - 3[a(x)]^2 + a(x). \end{aligned} \tag{1.57}$$

This establishes item (ii). The proof of Proposition 1.2.20 is thus complete. □

1.2.7.2 Integral of the standard logistic activation function

Lemma 1.2.21 (Primitive of the standard logistic activation function). *Let $\mathfrak{s}$ be the softplus activation function and let $\mathfrak{l}$ be the standard logistic activation function (cf. Definitions 1.2.11 and 1.2.18). Then it holds for all $x \in \mathbb{R}$ that*

$$\int_{-\infty}^{x} \mathfrak{l}(y)\,\mathrm{d}y = \int_{-\infty}^{x} \left(\frac{1}{1+e^{-y}}\right)\mathrm{d}y = \ln(1+\exp(x)) = \mathfrak{s}(x). \tag{1.58}$$

Proof of Lemma 1.2.21. Observe that (1.47) implies that for all $x \in \mathbb{R}$ it holds that

$$\mathfrak{s}'(x) = \left[\frac{1}{1+\exp(x)}\right]\exp(x) = \mathfrak{l}(x). \tag{1.59}$$

The fundamental theorem of calculus hence shows that for all $w, x \in \mathbb{R}$ with $w \leq x$ it holds that

$$\int_w^x \underbrace{\mathfrak{l}'(y)}_{\geq 0} \, \mathrm{d}y = \mathfrak{s}(x) - \mathfrak{s}(w). \tag{1.60}$$

Combining this with the fact that $\lim_{w \to -\infty} \mathfrak{s}(w) = 0$ establishes (1.58). The proof of Lemma 1.2.21 is thus complete. □

1.2.8 Swish activation

Definition 1.2.22 (Swish activation functions)**.** *Let $\beta \in \mathbb{R}$. Then we say that a is the swish activation function with parameter β if and only if it holds that $a \colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \frac{x}{1 + \exp(-\beta x)}. \tag{1.61}$$

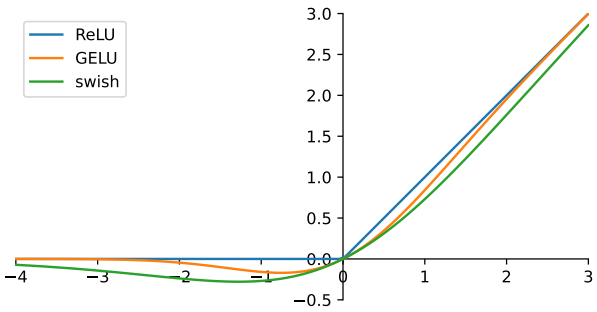

Figure 1.9 (plots/swish.pdf): A plot of the swish activation function with parameter 1, the GELU activation function, and the ReLU activation function

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-4,3), (-.5,3))

x = np.linspace(-4, 3, 100)

ax.plot(x, tf.keras.activations.relu(x), label='ReLU')
ax.plot(x, tf.keras.activations.gelu(x), label='GELU')
ax.plot(x, tf.keras.activations.swish(x), label='swish')
ax.legend()

```

```
15 plt.savefig("../../plots/swish.pdf", bbox_inches='tight')
```

Source code 1.7 (code/activation_functions/swish_plot.py): PYTHON code used to create Figure 1.9

Lemma 1.2.23 (Relation between swish activation functions and the logistic activation function)**.** *Let $\beta \in \mathbb{R}$, let $\mathfrak{s}$ be the swish activation function with parameter β, and let $\mathfrak{l}$ be the standard logistic activation function (cf. Definitions 1.2.18 and 1.2.22). Then it holds for all $x \in \mathbb{R}$ that*

$$\mathfrak{s}(x) = x\mathfrak{l}(\beta x). \tag{1.62}$$

Proof of Lemma 1.2.23. Observe that (1.61) and (1.53) establish (1.62). The proof of Lemma 1.2.23 is thus complete. □

Definition 1.2.24 (Multi-dimensional swish activation functions)**.** *Let $d \in \mathbb{N}$, $\beta \in \mathbb{R}$ and let a be the swish activation function with parameter β (cf. Definition 1.2.22). Then we say that A is the d-dimensional swish activation function with parameter β if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.9 Hyperbolic tangent activation

Definition 1.2.25 (Hyperbolic tangent activation function)**.** *We denote by* $\tanh\colon \mathbb{R} \to \mathbb{R}$ *the function which satisfies for all $x \in \mathbb{R}$ that*

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{1.63}$$

and we call tanh *the hyperbolic tangent activation function (we call* tanh *the hyperbolic tangent).*

```
1  import numpy as np
2  import tensorflow as tf
3  import matplotlib.pyplot as plt
4  import plot_util
5
6  ax = plot_util.setup_axis((-3,3), (-1.5,1.5))
7
8  x = np.linspace(-3, 3, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x+1, max_value=2)-1,
```

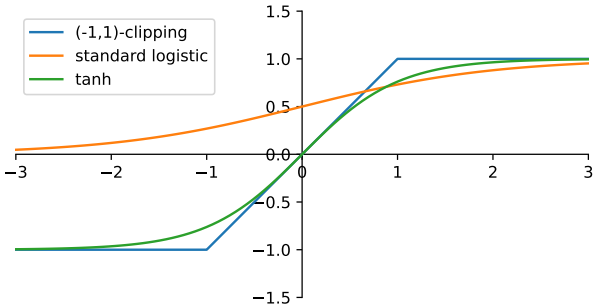

Figure 1.10 (plots/tanh.pdf): A plot of the hyperbolic tangent, the $(-1, 1)$-clipping activation function, and the standard logistic activation function

```
        label='(-1,1)-clipping')
ax.plot(x, tf.keras.activations.sigmoid(x),
        label='standard logistic')
ax.plot(x, tf.keras.activations.tanh(x), label='tanh')
ax.legend()

plt.savefig("../../plots/tanh.pdf", bbox_inches='tight')
```

Source code 1.8 (code/activation_functions/tanh_plot.py): PYTHON code used to create Figure 1.10

Definition 1.2.26 (Multi-dimensional hyperbolic tangent activation functions)**.** *Let $d \in \mathbb{N}$. Then we say that A is the d-dimensional hyperbolic tangent activation function if and only if $A = \mathfrak{M}_{\tanh,d}$ (cf. Definitions 1.2.1 and 1.2.25).*

Lemma 1.2.27. *Let a be the standard logistic activation function (cf. Definition 1.2.18). Then it holds for all $x \in \mathbb{R}$ that*

$$\tanh(x) = 2\,a(2x) - 1 \tag{1.64}$$

(cf. Definitions 1.2.18 and 1.2.25).

Proof of Lemma 1.2.27. Observe that (1.53) and (1.63) ensure that for all $x \in \mathbb{R}$ it holds

that

$$\begin{aligned} 2\,a(2x) - 1 &= 2\left(\frac{\exp(2x)}{\exp(2x)+1}\right) - 1 = \frac{2\exp(2x) - (\exp(2x)+1)}{\exp(2x)+1} \\ &= \frac{\exp(2x)-1}{\exp(2x)+1} = \frac{\exp(x)(\exp(x)-\exp(-x))}{\exp(x)(\exp(x)+\exp(-x))} \\ &= \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)} = \tanh(x). \end{aligned} \tag{1.65}$$

The proof of Lemma 1.2.27 is thus complete. □

Exercise 1.2.17. Let a be the standard logistic activation function (cf. Definition 1.2.18). Prove or disprove the following statement: There exists $L \in \{2, 3, \ldots\}$, $\mathfrak{d}, l_1, l_2, \ldots, l_{L-1} \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2\,l_1 + \left[\sum_{k=2}^{L-1} l_k(l_{k-1}+1)\right] + (l_{L-1}+1)$ such that for all $x \in \mathbb{R}$ it holds that

$$\left(\mathcal{N}^{\theta,1}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},\mathrm{id}_{\mathbb{R}}}\right)(x) = \tanh(x) \tag{1.66}$$

(cf. Definitions 1.1.3, 1.2.1, and 1.2.25).

1.2.10 Softsign activation

Definition 1.2.28 (Softsign activation function)**.** *We say that a is the softsign activation function if and only if it holds that $a\colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \frac{x}{|x|+1}. \tag{1.67}$$

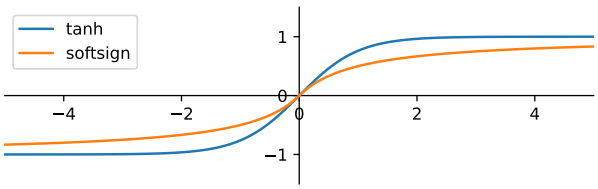

Figure 1.11 (plots/softsign.pdf): A plot of the softsign activation function and the hyperbolic tangent

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-5,5), (-1.5,1.5))
```

```
7
8  x = np.linspace(-5, 5, 100)
9
10 ax.plot(x, tf.keras.activations.tanh(x), label='tanh')
11 ax.plot(x, tf.keras.activations.softsign(x), label='softsign')
12 ax.legend()
13
14 plt.savefig("../../plots/softsign.pdf", bbox_inches='tight')
```

Source code 1.9 (code/activation_functions/softsign_plot.py): PYTHON code used to create Figure 1.11

Definition 1.2.29 (Multi-dimensional softsign activation functions)**.** *Let $d \in \mathbb{N}$ and let a be the softsign activation function (cf. Definition 1.2.28). Then we say that A is the d-dimensional softsign activation function if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.11 Leaky rectified linear unit (leaky ReLU) activation

Definition 1.2.30 (Leaky ReLU activation functions)**.** *Let $\gamma \in [0, \infty)$. Then we say that a is the leaky ReLU activation function with leak factor γ if and only if it holds that $a\colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \begin{cases} x & : x > 0 \\ \gamma x & : x \leq 0. \end{cases} \tag{1.68}$$

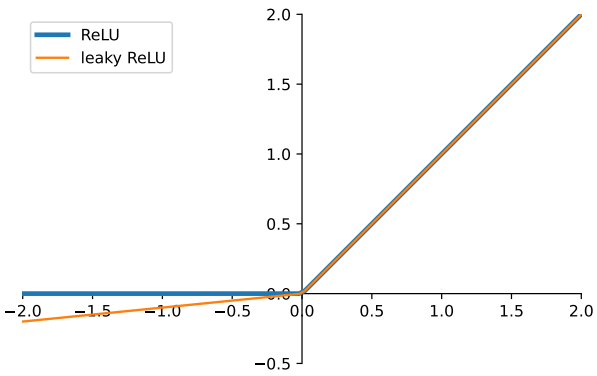

Figure 1.12 (plots/leaky_relu.pdf): A plot of the leaky ReLU activation function with leak factor $1/10$ and the ReLU activation function

```python
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-2,2), (-.5,2))

x = np.linspace(-2, 2, 100)

ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
ax.plot(x, tf.keras.activations.relu(x, alpha=0.1),
        label='leaky ReLU')
ax.legend()

plt.savefig("../../plots/leaky_relu.pdf", bbox_inches='tight')
```

Source code 1.10 (code/activation_functions/leaky_relu_plot.py): PYTHON code used to create Figure 1.12

Lemma 1.2.31. *Let $\gamma \in [0,1]$ and let $a\colon \mathbb{R} \to \mathbb{R}$ be a function. Then a is the leaky ReLU activation function with leak factor γ if and only if it holds for all $x \in \mathbb{R}$ that*

$$a(x) = \max\{x, \gamma x\} \tag{1.69}$$

(cf. Definition 1.2.30).

Proof of Lemma 1.2.31. Note that the fact that $\gamma \leq 1$ and (1.68) show (1.69). The proof of Lemma 1.2.31 is thus complete. □

Lemma 1.2.32. *Let $u, \beta \in \mathbb{R}$, $v \in (u, \infty)$, $\alpha \in (-\infty, 0]$, let a_1 be the softplus activation function, let a_2 be the GELU activation function, let a_3 be the standard logistic activation function, let a_4 be the swish activation function with parameter β, let a_5 be the softsign activation function, and let l be the leaky ReLU activation function with leaky parameter γ (cf. Definitions 1.2.11, 1.2.15, 1.2.18, 1.2.22, 1.2.28, and 1.2.30). Then*

(i) it holds for all $f \in \{\mathfrak{r}, \mathfrak{c}_{u,v}, \tanh, a_1, a_2, \ldots, a_5\}$ that $\limsup_{x \to -\infty} |f'(x)| = 0$ and

(ii) it holds that $\lim_{x \to -\infty} l'(x) = \gamma$

(cf. Definitions 1.2.4, 1.2.9, and 1.2.25).

Proof of Lemma 1.2.32. Note that (1.26), (1.45), (1.47), (1.52), (1.53), (1.61), (1.63), and (1.67) prove item (i). Observe that (1.68) establishes item (ii). The proof of Lemma 1.2.32 is thus complete. □

Definition 1.2.33 (Multi-dimensional leaky ReLU activation functions)**.** *Let $d \in \mathbb{N}$, $\gamma \in [0, \infty)$ and let a be the leaky ReLU activation function with leak factor γ (cf. Definition 1.2.30). Then we say that A is the d-dimensional leaky ReLU activation function with leak factor γ if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.12 Exponential linear unit (ELU) activation

Another popular activation function is the so-called *exponential linear unit* (ELU) activation function which has been introduced in Clevert et al. [85]. This activation function is the subject of the next notion.

Definition 1.2.34 (ELU activation functions)**.** *Let $\gamma \in (-\infty, 0]$. Then we say that a is the ELU activation function with asymptotic γ if and only if it holds that $a \colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \begin{cases} x & : x > 0 \\ \gamma(1 - \exp(x)) & : x \leq 0. \end{cases} \tag{1.70}$$

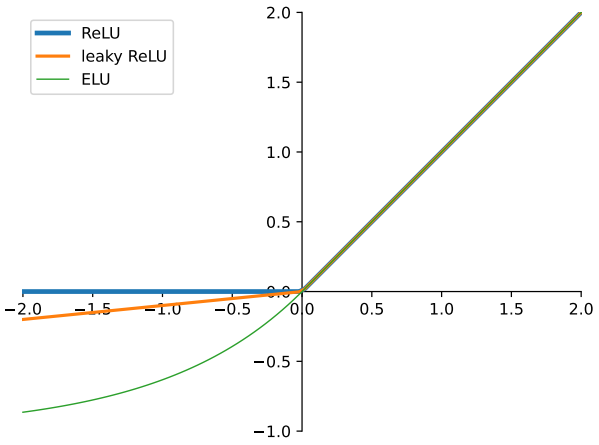

Figure 1.13 (plots/elu.pdf): A plot of the ELU activation function with asymptotic -1, the leaky ReLU activation function with leak factor $1/10$, and the ReLU activation function

```
1 import numpy as np
2 import tensorflow as tf
3 import matplotlib.pyplot as plt
4 import plot_util
5
```

```
ax = plot_util.setup_axis((-2,2), (-1,2))

x = np.linspace(-2, 2, 100)

ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
ax.plot(x, tf.keras.activations.relu(x, alpha=0.1), linewidth=2,
    label='leaky ReLU')
ax.plot(x, tf.keras.activations.elu(x), linewidth=0.9, label='ELU')
ax.legend()

plt.savefig("../../plots/elu.pdf", bbox_inches='tight')
```

Source code 1.11 (code/activation_functions/elu_plot.py): PYTHON code used to create Figure 1.13

Lemma 1.2.35. *Let $\gamma \in (-\infty, 0]$ and let a be the ELU activation function with asymptotic γ (cf. Definition 1.2.34). Then*

$$\limsup_{x \to -\infty} a(x) = \liminf_{x \to -\infty} a(x) = \gamma. \tag{1.71}$$

Proof of Lemma 1.2.35. Observe that (1.70) implies (1.71). The proof of Lemma 1.2.35 is thus complete. □

Definition 1.2.36 (Multi-dimensional ELU activation functions)**.** *Let $d \in \mathbb{N}$, $\gamma \in (-\infty, 0]$ and let a be the ELU activation function with asymptotic γ (cf. Definition 1.2.34). Then we say that A is the d-dimensional ELU activation function with asymptotic γ if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.13 Rectified power unit (RePU) activation

Another popular activation function is the so-called *rectified power unit* (RePU) activation function. This concept is the subject of the next notion.

Definition 1.2.37 (RePU activation functions)**.** *Let $p \in \mathbb{N}$. Then we say that a is the RePU activation function with power p if and only if it holds that $a \colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = (\max\{x, 0\})^p. \tag{1.72}$$

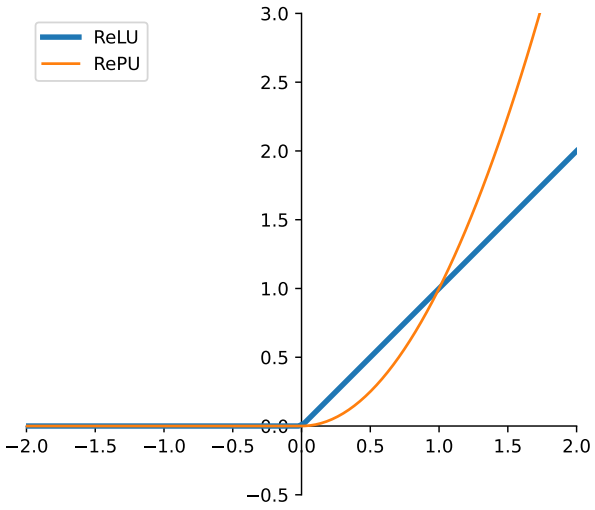

Figure 1.14 (plots/repu.pdf): A plot of the RePU activation function with power 2 and the ReLU activation function

```python
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-2,2), (-.5,3))
ax.set_ylim(-.5, 3)

x = np.linspace(-2, 2, 100)

ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
ax.plot(x, tf.keras.activations.relu(x)**2, label='RePU')
ax.legend()

plt.savefig("../../plots/repu.pdf", bbox_inches='tight')
```

Source code 1.12 (code/activation_functions/repu_plot.py): PYTHON code used to create Figure 1.14

Definition 1.2.38 (Multi-dimensional RePU activation functions)**.** *Let $d, p \in \mathbb{N}$ and let a be the RePU activation function with power p (cf. Definition 1.2.37). Then we say that A is the d-dimensional RePU activation function with power p if and only if it holds that $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.14 Sine activation

The sine function has been proposed as activation function in Sitzmann et al. [394]. This is formulated in the next notion.

Definition 1.2.39 (Sine activation function)**.** *We say that a is the sine activation function if and only if it holds that $a\colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \sin(x). \tag{1.73}$$

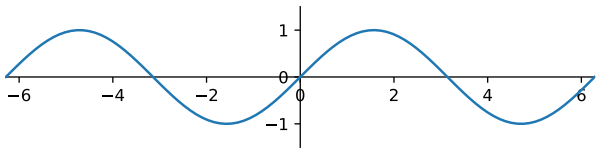

Figure 1.15 (plots/sine.pdf): A plot of the sine activation function

```
1  import numpy as np
2  import tensorflow as tf
3  import matplotlib.pyplot as plt
4  import plot_util
5
6  ax = plot_util.setup_axis((-2*np.pi,2*np.pi), (-1.5,1.5))
7
8  x = np.linspace(-2*np.pi, 2*np.pi, 100)
9
10 ax.plot(x, np.sin(x))
11
12 plt.savefig("../../plots/sine.pdf", bbox_inches='tight')
```

Source code 1.13 (code/activation_functions/sine_plot.py): PYTHON code used to create Figure 1.15

Definition 1.2.40 (Multi-dimensional sine activation functions)**.** *Let $d \in \mathbb{N}$ and let a be the sine activation function (cf. Definition 1.2.39). Then we say that A is the d-dimensional sine activation function if and only if it holds that $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.15 Heaviside activation

Definition 1.2.41 (Heaviside activation function)**.** *We say that a is the Heaviside activation function (we say that a is the Heaviside step function, we say that a is the unit step function) if and only if it holds that $a\colon \mathbb{R} \to \mathbb{R}$ is the function from $\mathbb{R}$ to $\mathbb{R}$ which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \mathbb{1}_{[0,\infty)}(x) = \begin{cases} 1 & : x \geq 0 \\ 0 & : x < 0. \end{cases} \tag{1.74}$$

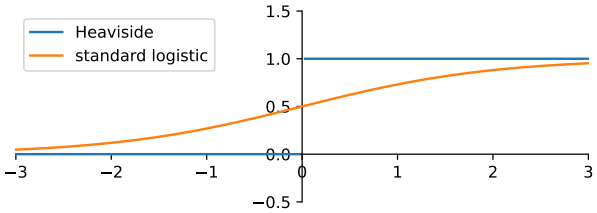

Figure 1.16 (plots/heaviside.pdf): A plot of the Heaviside activation function and the standard logistic activation function

```python
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-3,3), (-.5,1.5))

x = np.linspace(-3, 3, 100)

ax.plot(x[0:50], [0]*50, 'C0')
ax.plot(x[50:100], [1]*50, 'C0', label='Heaviside')
ax.plot(x, tf.keras.activations.sigmoid(x), 'C1',
        label='standard logistic')
ax.legend()

plt.savefig("../../plots/heaviside.pdf", bbox_inches='tight')
```

Source code 1.14 (code/activation_functions/heaviside_plot.py): PYTHON code used to create Figure 1.16

Definition 1.2.42 (Multi-dimensional Heaviside activation functions)**.** *Let $d \in \mathbb{N}$ and let a be the Heaviside activation function (cf. Definition 1.2.41). Then we say that A is the d-dimensional Heaviside activation function (we say that A is the d-dimensional Heaviside step function, we say that A is the d-dimensional unit step function) if and*

only if it holds that $A = \mathfrak{M}_{a,d}$ *(cf. Definition 1.2.1).*

1.2.16 Softmax activation

Definition 1.2.43 (Softmax activation functions)**.** *Let* $d \in \mathbb{N}$*. Then we say that* A *is the* d*-dimensional softmax activation function if and only if it holds that* $A \colon \mathbb{R}^d \to \mathbb{R}^d$ *is the function from* $\mathbb{R}^d$ *to* $\mathbb{R}^d$ *which satisfies for all* $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ *that*

$$A(x) = \left(\frac{\exp(x_1)}{\left(\sum_{i=1}^{d} \exp(x_i)\right)}, \frac{\exp(x_2)}{\left(\sum_{i=1}^{d} \exp(x_i)\right)}, \ldots, \frac{\exp(x_d)}{\left(\sum_{i=1}^{d} \exp(x_i)\right)} \right). \tag{1.75}$$

Lemma 1.2.44. *Let* $d \in \mathbb{N}$ *and let* $A = (A_1, \ldots, A_d)$ *be the* d*-dimensional softmax activation function (cf. Definition 1.2.43). Then*

(i) it holds for all $x \in \mathbb{R}^d$*,* $k \in \{1, 2, \ldots, d\}$ *that* $A_k(x) \in (0, 1]$ *and*

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

$$\sum_{k=1}^{d} A_k(x) = 1. \tag{1.76}$$

tum

(cf. Definition 1.2.43).

Proof of Lemma 1.2.44. Observe that (1.75) demonstrates that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ it holds that

$$\sum_{k=1}^{d} A_k(x) = \sum_{k=1}^{d} \frac{\exp(x_k)}{\left(\sum_{i=1}^{d} \exp(x_i)\right)} = \frac{\sum_{k=1}^{d} \exp(x_k)}{\sum_{i=1}^{d} \exp(x_i)} = 1. \tag{1.77}$$

The proof of Lemma 1.2.44 is thus complete. □

1.3 Fully-connected feedforward ANNs (structured description)

In this section we present an alternative way to describe the fully-connected feedforward ANNs introduced in Section 1.1 above. Roughly speaking, in Section 1.1 above we defined a *vectorized description* of fully-connected feedforward ANNs in the sense that the trainable parameters of a fully-connected feedforward 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 fully-connected feedforward ANNs in which the trainable parameters of

a fully-connected feedforward ANN are represented by a tuple of matrix-vector pairs corresponding to the weight matrices and bias vectors of the fully-connected feedforward ANNs (cf. Definitions 1.3.1 and 1.3.4 below).

1.3.1 Structured description of fully-connected feedforward ANNs

Definition 1.3.1 (Structured description of fully-connected feedforward ANNs)**.** *We denote by* $\mathbf{N}$ *the set given by*

$$\mathbf{N} = \bigcup_{L\in\mathbb{N}} \bigcup_{l_0,l_1,\ldots,l_L\in\mathbb{N}} \left(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k\times l_{k-1}} \times \mathbb{R}^{l_k}) \right), \tag{1.78}$$

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

$$\mathcal{P}(\Phi) = \textstyle\sum_{k=1}^{L} l_k(l_{k-1}+1), \quad \mathcal{L}(\Phi) = L, \quad \mathcal{I}(\Phi) = l_0, \quad \mathcal{O}(\Phi) = l_L, \quad \text{and} \quad \mathcal{H}(\Phi) = L-1, \tag{1.79}$$

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

$$\mathbb{D}_n(\Phi) = \begin{cases} l_n & : n \leq L \\ 0 & : n > L, \end{cases} \tag{1.80}$$

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

$$\mathcal{D}(\Phi) = (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \ldots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \tag{1.81}$$

and for every $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$, $\Phi = ((W_1, B_1), \ldots, (W_L, B_L)) \in \left(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k\times l_{k-1}} \times \mathbb{R}^{l_k}) \right) \subseteq \mathbf{N}$, $n \in \{1, 2, \ldots, L\}$ *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*

$$\mathcal{W}_{n,\Phi} = W_n \qquad \text{and} \qquad \mathcal{B}_{n,\Phi} = B_n. \tag{1.82}$$

Definition 1.3.2 (Fully-connected feedforward ANNs)**.** *We say that* Φ *is a fully-connected feedforward ANN if and only if it holds that*

$$\Phi \in \mathbf{N} \tag{1.83}$$

(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

$$\mathcal{I}(\Phi) = \mathbb{D}_0(\Phi) \qquad \text{and} \qquad \mathcal{O}(\Phi) = \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi), \tag{1.84}$$

and

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

$$\mathcal{W}_{n,\Phi} \in \mathbb{R}^{\mathbb{D}_n(\Phi) \times \mathbb{D}_{n-1}(\Phi)} \qquad \text{and} \qquad \mathcal{B}_{n,\Phi} \in \mathbb{R}^{\mathbb{D}_n(\Phi)}. \tag{1.85}$$

.

Proof of Lemma 1.3.3. Note that the assumption that

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

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

$$\Phi \in \left(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right). \tag{1.86}$$

Observe that (1.86), (1.79), and (1.80) imply that

$$\mathcal{L}(\Phi) = L, \qquad \mathcal{I}(\Phi) = l_0 = \mathbb{D}_0(\Phi), \qquad \text{and} \qquad \mathcal{O}(\Phi) = l_L = \mathbb{D}_L(\Phi). \tag{1.87}$$

This shows that

$$\mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_L) \in \mathbb{N}^{L+1} = \mathbb{N}^{\mathcal{L}(\Phi)+1}. \tag{1.88}$$

Next note that (1.86), (1.80), and (1.82) ensure that for all $n \in \{1, 2, \ldots, \mathcal{L}(\Phi)\}$ it holds that

$$\mathcal{W}_{n,\Phi} \in \mathbb{R}^{l_n \times l_{n-1}} = \mathbb{R}^{\mathbb{D}_n(\Phi) \times \mathbb{D}_{n-1}(\Phi)} \qquad \text{and} \qquad \mathcal{B}_{n,\Phi} \in \mathbb{R}^{l_n} = \mathbb{R}^{\mathbb{D}_n(\Phi)}. \tag{1.89}$$

The proof of Lemma 1.3.3 is thus complete. □

1.3.2 Realizations of fully-connected feedforward ANNs

Definition 1.3.4 (Realizations of fully-connected feedforward ANNs)**.** *Let $\Phi \in \mathbf{N}$ and let $a \colon \mathbb{R} \to \mathbb{R}$ be a function (cf. Definition 1.3.1). Then we denote by*

$$\mathcal{R}_a^{\mathbf{N}}(\Phi) \colon \mathbb{R}^{\mathcal{I}(\Phi)} \to \mathbb{R}^{\mathcal{O}(\Phi)} \tag{1.90}$$

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

$$\forall\, k \in \{1, 2, \ldots, \mathcal{L}(\Phi)\} \colon x_k = \mathfrak{M}_{a \mathbb{1}_{(0, \mathcal{L}(\Phi))}(k) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{\mathcal{L}(\Phi)\}}(k), \mathbb{D}_k(\Phi)}(\mathcal{W}_{k,\Phi} x_{k-1} + \mathcal{B}_{k,\Phi}) \tag{1.91}$$

that

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi))(x_0) = x_{\mathcal{L}(\Phi)} \tag{1.92}$$

and we call $\mathcal{R}_a^{\mathbf{N}}(\Phi)$ the realization function of the fully-connected feedforward ANN Φ with activation function a (we call $\mathcal{R}_a^{\mathbf{N}}(\Phi)$ the realization of the fully-connected feedforward ANN Φ with activation 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 $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$, let $a\colon \mathbb{R} \to \mathbb{R}$ be a function, and consider the function

$$\mathscr{f}\colon \left(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right) \times \mathbb{R}^{l_0} \to \mathbb{R}^{l_L} \tag{1.93}$$

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

$$\mathscr{f}(\Phi, x) = \mathcal{R}_a^{\mathbf{N}}(\Phi)(x) \tag{1.94}$$

(cf. Definition 1.3.4). In this context, the function $\mathscr{f}$ itself is sometimes referred to as an ANN.

Exercise 1.3.1. Let

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3)) \in (\mathbb{R}^{2\times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R}^1) \tag{1.95}$$

satisfy

$$W_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} -1 & 2 \\ 3 & -4 \\ -5 & 6 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \tag{1.96}$$

$$W_3 = \begin{pmatrix} -1 & 1 & -1 \end{pmatrix}, \qquad \text{and} \qquad B_3 = \begin{pmatrix} -4 \end{pmatrix}. \tag{1.97}$$

Prove or disprove the following statement: It holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(-1) = 0 \tag{1.98}$$

(cf. Definitions 1.2.4 and 1.3.4).

Exercise 1.3.2. Let a be the standard logistic activation function (cf. Definition 1.2.18). Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that

$$\mathcal{R}^{\mathbf{N}}_{\tanh}(\Phi) = a \tag{1.99}$$

(cf. Definitions 1.2.25, 1.3.1, and 1.3.4).

```python
import torch
import torch.nn as nn
import torch.nn.functional as F


# To define a neural network, we define a class that inherits from
# torch.nn.Module
class FullyConnectedANN(nn.Module):
    def __init__(self):
        super().__init__()
        # In the constructor, we define the weights and biases.
        # Wrapping the tensors in torch.nn.Parameter objects tells
        # PyTorch that these are parameters that should be
        # optimized during training.
        self.W1 = nn.Parameter(
            torch.Tensor([[1, 0], [0, -1], [-2, 2]])
        )
        self.B1 = nn.Parameter(torch.Tensor([0, 2, -1]))
        self.W2 = nn.Parameter(torch.Tensor([[1, -2, 3]]))
        self.B2 = nn.Parameter(torch.Tensor([1]))

    # The realization function of the network
    def forward(self, x0):
        x1 = F.relu(self.W1 @ x0 + self.B1)
        x2 = self.W2 @ x1 + self.B2
        return x2


model = FullyConnectedANN()

x0 = torch.Tensor([1, 2])
# Print the output of the realization function for input x0
print(model.forward(x0))

# As a consequence of inheriting from torch.nn.Module we can just
# "call" the model itself (which will call the forward method
# implicitly)
print(model(x0))

# Wrapping a tensor in a Parameter object and assigning it to an
# instance variable of the Module makes PyTorch register it as a
# parameter. We can access all parameters via the parameters
```

```
# method.
for p in model.parameters():
    print(p)
```

Source code 1.15 (code/fc-ann-manual.py): PYTHON code for implementing a fully-connected feedforward ANN in PYTORCH. The model created here represents the fully-connected feedforward ANN $\left(\left(\begin{pmatrix} 1 & 0 \\ 0 & -1 \\ -2 & 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}\right), \left(\begin{pmatrix} 1 & -2 & 3 \end{pmatrix}, \begin{pmatrix} 1 \end{pmatrix}\right)\right) \in (\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R}^1) \subseteq \mathbf{N}$ using the ReLU activation function after the hidden layer.

```
import torch
import torch.nn as nn


class FullyConnectedANN(nn.Module):
    def __init__(self):
        super().__init__()
        # Define the layers of the network in terms of Modules.
        # nn.Linear(3, 20) represents an affine function defined
        # by a 20x3 weight matrix and a 20-dimensional bias vector.
        self.affine1 = nn.Linear(3, 20)
        # The torch.nn.ReLU class simply wraps the
        # torch.nn.functional.relu function as a Module.
        self.activation1 = nn.ReLU()
        self.affine2 = nn.Linear(20, 30)
        self.activation2 = nn.ReLU()
        self.affine3 = nn.Linear(30, 1)

    def forward(self, x0):
        x1 = self.activation1(self.affine1(x0))
        x2 = self.activation2(self.affine2(x1))
        x3 = self.affine3(x2)
        return x3


model = FullyConnectedANN()

x0 = torch.Tensor([1, 2, 3])
print(model(x0))

# Assigning a Module to an instance variable of a Module registers
# all of the former's parameters as parameters of the latter
for p in model.parameters():
    print(p)
```

Source code 1.16 (code/fc-ann.py): PYTHON code for implementing a fully-connected feedforward ANN in PYTORCH. The model implemented here represents a fully-connected feedforward ANN with two hidden layers, 3 neurons in the input layer, 20 neurons in the first hidden layer, 30 neurons in the second hidden layer, and 1 neuron in the output layer. Unlike Source code 1.15, this code uses the `torch.nn.Linear` class to represent the affine transformations.

```
import torch
import torch.nn as nn

# A Module whose forward method is simply a composition of Modules
# can be represented using the torch.nn.Sequential class
model = nn.Sequential(
    nn.Linear(3, 20),
    nn.ReLU(),
    nn.Linear(20, 30),
    nn.ReLU(),
    nn.Linear(30, 1),
)

# Prints a summary of the model architecture
print(model)

x0 = torch.Tensor([1, 2, 3])
print(model(x0))
```

Source code 1.17 (code/fc-ann2.py): PYTHON code for creating a fully-connected feedforward ANN in PYTORCH. This creates the same model as Source code 1.16 but uses the `torch.nn.Sequential` class instead of defining a new subclass of `torch.nn.Module`.

1.3.3 On the connection to the vectorized description

Definition 1.3.6 (Transformation from the structured to the vectorized description of fully-connected feedforward ANNs)**.** *We denote by* $\mathcal{T} \colon \mathbf{N} \to \left(\bigcup_{d \in \mathbb{N}} \mathbb{R}^d\right)$ *the function which satisfies for all* $\Phi \in \mathbf{N}$, $k \in \{1, 2, \ldots, \mathcal{L}(\Phi)\}$, $d \in \mathbb{N}$, $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ *with*

$\mathcal{T}(\Phi) = \theta$ *that*

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

$$\begin{pmatrix} \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+l_{k-1}} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+l_{k-1}+1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+l_{k-1}+2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+2l_{k-1}} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+2l_{k-1}+1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+2l_{k-1}+2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+3l_{k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+(l_k-1)l_{k-1}+1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+(l_k-1)l_{k-1}+2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1))+l_k l_{k-1}} \end{pmatrix} \tag{1.100}$$

(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*

$$\Phi = \left(\left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix} \right), \left(\begin{pmatrix} 13 & 14 & 15 \\ 16 & 17 & 18 \end{pmatrix}, \begin{pmatrix} 19 \\ 20 \end{pmatrix} \right) \right). \tag{1.101}$$

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

Proof for Example 1.3.7. Observe that (1.100) establishes (1.101). The proof for Example 1.3.7 is thus complete. □

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

$$\begin{aligned} &\mathcal{T}\big(((W, B))\big) \\ &= \big(W_{1,1}, W_{1,2}, \ldots, W_{1,b}, W_{2,1}, W_{2,2}, \ldots, W_{2,b}, \ldots, W_{a,1}, W_{a,2}, \ldots, W_{a,b}, B_1, B_2, \ldots, B_a\big) \end{aligned} \tag{1.102}$$

(cf. Definition 1.3.6).

Proof of Lemma 1.3.8. Observe that (1.100) establishes (1.102). The proof of Lemma 1.3.8 is thus complete. □

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

$$\begin{aligned}
&\mathcal{T}\Big(\big((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)\big)\Big) \\
&= \Big(W_{1,1,1}, W_{1,1,2}, \ldots, W_{1,1,l_0}, \ldots, W_{1,l_1,1}, W_{1,l_1,2}, \ldots, W_{1,l_1,l_0}, B_{1,1}, B_{1,2}, \ldots, B_{1,l_1}, \\
&\quad W_{2,1,1}, W_{2,1,2}, \ldots, W_{2,1,l_1}, \ldots, W_{2,l_2,1}, W_{2,l_2,2}, \ldots, W_{2,l_2,l_1}, B_{2,1}, B_{2,2}, \ldots, B_{2,l_2}, \\
&\quad \ldots, \\
&\quad W_{L,1,1}, W_{L,1,2}, \ldots, W_{L,1,l_{L-1}}, \ldots, W_{L,l_L,1}, W_{L,l_L,2}, \ldots, W_{L,l_L,l_{L-1}}, B_{L,1}, B_{L,2}, \ldots, B_{L,l_L}\Big)
\end{aligned} \tag{1.103}$$

(cf. Definition 1.3.6).

Proof of Lemma 1.3.9. Note that (1.100) implies (1.103). The proof of Lemma 1.3.9 is thus complete. □

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

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

Exercise 1.3.5. Prove or disprove the following statement: The function $\mathcal{T}$ is bijective (cf. Definition 1.3.6).

Proposition 1.3.10. *Let $a\colon \mathbb{R} \to \mathbb{R}$ be a function and let $\Phi \in \mathbf{N}$. (cf. Definition 1.3.1). Then*

$$\mathcal{R}_a^{\mathbf{N}}(\Phi) = \begin{cases} \mathcal{N}^{\mathcal{T}(\Phi),\mathcal{I}(\Phi)}_{\mathrm{id}_{\mathbb{R}^{\mathcal{O}(\Phi)}}} & : \mathcal{H}(\Phi) = 0 \\ \mathcal{N}^{\mathcal{T}(\Phi),\mathcal{I}(\Phi)}_{\mathfrak{M}_{a,\mathbb{D}_1(\Phi)},\mathfrak{M}_{a,\mathbb{D}_2(\Phi)},\ldots,\mathfrak{M}_{a,\mathbb{D}_{\mathcal{H}(\Phi)}(\Phi)},\mathrm{id}_{\mathbb{R}^{\mathcal{O}(\Phi)}}} & : \mathcal{H}(\Phi) > 0 \end{cases} \tag{1.104}$$

(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 $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$ satisfy

$$\mathcal{L}(\Phi) = L \qquad \text{and} \qquad \mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_L). \tag{1.105}$$

Note that (1.100) shows that for all $k \in \{1, 2, \ldots, L\}$, $x \in \mathbb{R}^{l_{k-1}}$ it holds that

$$\mathcal{W}_{k,\Phi} x + \mathcal{B}_{k,\Phi} = \big(\mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{k-1} l_i(l_{i-1}+1)}_{l_k,l_{k-1}}\big)(x) \tag{1.106}$$

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

$$x_{L-1} = \begin{cases} x_0 & : L = 1 \\ \big(\mathfrak{M}_{a,l_{L-1}} \circ \mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{L-2} l_i(l_{i-1}+1)}_{l_{L-1},l_{L-2}} \\ \quad \circ \mathfrak{M}_{a,l_{L-2}} \circ \mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{L-3} l_i(l_{i-1}+1)}_{l_{L-2},l_{L-3}} \circ \ldots \circ \mathfrak{M}_{a,l_1} \circ \mathcal{A}^{\mathcal{T}(\Phi),0}_{l_1,l_0}\big)(x_0) & : L > 1 \end{cases} \tag{1.107}$$

(cf. Definition 1.2.1). This, (1.106), (1.5), and (1.92) prove that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}, \ldots, x_L \in \mathbb{R}^{l_L}$ with $\forall\, k \in \{1, 2, \ldots, L\}\colon x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k)+\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(k),l_k}(\mathcal{W}_{k,\Phi} x_{k-1} + \mathcal{B}_{k,\Phi})$ it holds that

$$\begin{aligned} \big(\mathcal{R}^{\mathbf{N}}_a(\Phi)\big)(x_0) = x_L &= \mathcal{W}_{L,\Phi} x_{L-1} + \mathcal{B}_{L,\Phi} = \big(\mathcal{A}^{\mathcal{T}(\Phi),\sum_{i=1}^{L-1} l_i(l_{i-1}+1)}_{l_L,l_{L-1}}\big)(x_{L-1}) \\ &= \begin{cases} \big(\mathcal{N}^{\mathcal{T}(\Phi),l_0}_{\mathrm{id}_{\mathbb{R}^{l_L}}}\big)(x_0) & : L = 1 \\ \big(\mathcal{N}^{\mathcal{T}(\Phi),l_0}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},\mathrm{id}_{\mathbb{R}^{l_L}}}\big)(x_0) & : L > 1 \end{cases} \end{aligned} \tag{1.108}$$

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

1.4 Convolutional ANNs (CNNs)

In this section we review CNNs, which are ANNs designed to process data with a spatial structure. In a broad sense, CNNs can be thought of as any ANNs involving a convolution operation (cf, for instance, Definition 1.4.1 below). Roughly speaking, convolutional operations allow CNNs to exploit spatial invariance of data by performing the same operations across different regions of an input data point. In principle, such convolution operations can be employed in combinations with other ANN architecture elements, such as fully-connected layers (cf., for example, Sections 1.1 and 1.3 above), residual layers (cf., for instance, Section 1.5 below), and recurrent structures (cf., for example, Section 1.6 below). However, for simplicity we introduce in this section in all mathematical details feedforward CNNs only involving convolutional layers based on the discrete convolution operation without *padding* (sometimes called *valid padding*) in Definition 1.4.1 (see Definitions 1.4.2 and 1.4.5 below). We refer, for instance, to [4, Section 12.5], [36, Sectino 1.6.1], [61, Chapter 16], [64, Section 4.2], [170, Chapter 9], and [279] for other introductions on CNNs.

CNNs were introduced in LeCun et al. [272] for *computer vision* (CV) applications. The first successful modern CNN architecture is widely considered to be the *AlexNet* architecture proposed in Krizhevsky et al. [267]. A few other very successful early CNN architecures for CV include [158, 198, 214, 293, 304, 385, 392, 404]. While CV is by far the most popular domain of application for CNNs, CNNs have also been employed successfully in several other areas. In particular, we refer, for example, to [115, 149, 255, 444, 448, 451] for applications of CNNs to *natural language processing* (NLP), we refer, for instance, to [1, 60, 80, 373, 410]

for applications of CNNs to audio processing, and we refer, for example, to [47, 110, 246, 362, 422, 454] for applications of CNNs to time series analysis. Finally, for approximation results for feedforward CNNs we refer, for instance, to Petersen & Voigtländer [348] and the references therein.

1.4.1 Discrete convolutions

Definition 1.4.1 (Discrete convolutions). *Let $T \in \mathbb{N}$, $a_1, a_2, \ldots, a_T, w_1, w_2, \ldots, w_T, \mathfrak{d}_1, \mathfrak{d}_2, \ldots, \mathfrak{d}_T \in \mathbb{N}$ and let $A = (A_{i_1,i_2,\ldots,i_T})_{(i_1,i_2,\ldots,i_T)\in(\times_{t=1}^{T}\{1,2,\ldots,a_t\})} \in \mathbb{R}^{a_1\times a_2\times\ldots\times a_T}$, $W = (W_{i_1,i_2,\ldots,i_T})_{(i_1,i_2,\ldots,i_T)\in(\times_{t=1}^{T}\{1,2,\ldots,w_t\})} \in \mathbb{R}^{w_1\times w_2\times\ldots\times w_T}$ satisfy for all $t \in \{1,2,\ldots,T\}$ that*

$$\mathfrak{d}_t = a_t - w_t + 1. \tag{1.109}$$

Then we denote by $A * W = ((A*W)_{i_1,i_2,\ldots,i_T})_{(i_1,i_2,\ldots,i_T)\in(\times_{t=1}^{T}\{1,2,\ldots,\mathfrak{d}_t\})} \in \mathbb{R}^{\mathfrak{d}_1\times\mathfrak{d}_2\times\ldots\times\mathfrak{d}_T}$ the tensor which satisfies for all $i_1 \in \{1,2,\ldots,\mathfrak{d}_1\}$, $i_2 \in \{1,2,\ldots,\mathfrak{d}_2\}$, $\ldots$, $i_T \in \{1,2,\ldots,\mathfrak{d}_T\}$ that

$$(A*W)_{i_1,i_2,\ldots,i_T} = \sum_{r_1=1}^{w_1}\sum_{r_2=1}^{w_2}\cdots\sum_{r_T=1}^{w_T} A_{i_1-1+r_1,i_2-1+r_2,\ldots,i_T-1+r_T} W_{r_1,r_2,\ldots,r_T}. \tag{1.110}$$

1.4.2 Structured description of feedforward CNNs

Definition 1.4.2 (Structured description of feedforward CNNs). *We denote by $\mathbf{C}$ the set given by*

$$\mathbf{C} = \bigcup_{T,L\in\mathbb{N}}\bigcup_{l_0,l_1,\ldots,l_L\in\mathbb{N}}\bigcup_{(c_{k,t})_{(k,t)\in\{1,2,\ldots,L\}\times\{1,2,\ldots,T\}}\subseteq\mathbb{N}}\left(\bigtimes_{k=1}^{L}\left((\mathbb{R}^{c_{k,1}\times c_{k,2}\times\ldots\times c_{k,T}})^{l_k\times l_{k-1}}\times\mathbb{R}^{l_k}\right)\right). \tag{1.111}$$

Definition 1.4.3 (Feedforward CNNs). *We say that Φ is a feedforward CNN if and only if it holds that*

$$\Phi \in \mathbf{C} \tag{1.112}$$

(cf. Definition 1.4.2).

1.4.3 Realizations of feedforward CNNs

Definition 1.4.4 (One tensor). *Let $T \in \mathbb{N}$, $d_1, d_2, \dots, d_T \in \mathbb{N}$. Then we denote by $\mathbf{I}^{d_1,d_2,\dots,d_T} = (\mathbf{I}^{d_1,d_2,\dots,d_T}_{i_1,i_2,\dots,i_T})_{(i_1,i_2,\dots,i_T)\in(\bigtimes_{t=1}^{T}\{1,2,\dots,d_t\})} \in \mathbb{R}^{d_1\times d_2\times\dots\times d_T}$ the tensor which satisfies for all $i_1 \in \{1,2,\dots,d_1\}$, $i_2 \in \{1,2,\dots,d_2\}$, $\dots$, $i_T \in \{1,2,\dots,d_T\}$ that*

$$\mathbf{I}^{d_1,d_2,\dots,d_T}_{i_1,i_2,\dots,i_T} = 1. \tag{1.113}$$

Definition 1.4.5 (Realizations associated to feedforward CNNs). *Let $T, L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, let $(c_{k,t})_{(k,t)\in\{1,2,\dots,L\}\times\{1,2,\dots,T\}} \subseteq \mathbb{N}$, let $\Phi = (((W_{k,n,m})_{(n,m)\in\{1,2,\dots,l_k\}\times\{1,2,\dots,l_{k-1}\}}, (B_{k,n})_{n\in\{1,2,\dots,l_k\}}))_{k\in\{1,2,\dots,L\}} \in \bigtimes_{k=1}^{L}((\mathbb{R}^{c_{k,1}\times c_{k,2}\times\dots\times c_{k,T}})^{l_k\times l_{k-1}}\times\mathbb{R}^{l_k}) \subseteq \mathbf{C}$, and let $a\colon\mathbb{R}\to\mathbb{R}$ be a function. Then we denote by*

$$\mathcal{R}^{\mathbf{C}}_a(\Phi)\colon \left(\bigcup_{\substack{d_1,d_2,\dots,d_T\in\mathbb{N}\\ \forall\, t\in\{1,2,\dots,T\}\colon d_t-\sum_{k=1}^{L}(c_{k,t}-1)\geq 1}} (\mathbb{R}^{d_1\times d_2\times\dots\times d_T})^{l_0}\right) \to \left(\bigcup_{d_1,d_2,\dots,d_T\in\mathbb{N}} (\mathbb{R}^{d_1\times d_2\times\dots\times d_T})^{l_L}\right) \tag{1.114}$$

the function which satisfies for all $(\mathfrak{d}_{k,t})_{(k,t)\in\{0,1,\dots,L\}\times\{1,2,\dots,T\}} \subseteq \mathbb{N}$, $x_0 = (x_{0,1},\dots,x_{0,l_0}) \in (\mathbb{R}^{\mathfrak{d}_{0,1}\times\mathfrak{d}_{0,2}\times\dots\times\mathfrak{d}_{0,T}})^{l_0}$, $x_1 = (x_{1,1},\dots,x_{1,l_1}) \in (\mathbb{R}^{\mathfrak{d}_{1,1}\times\mathfrak{d}_{1,2}\times\dots\times\mathfrak{d}_{1,T}})^{l_1}$, $\dots$, $x_L = (x_{L,1},\dots,x_{L,l_L}) \in (\mathbb{R}^{\mathfrak{d}_{L,1}\times\mathfrak{d}_{L,2}\times\dots\times\mathfrak{d}_{L,T}})^{l_L}$ with

$$\forall\, k\in\{1,2,\dots,L\},\ t\in\{1,2,\dots,T\}\colon \mathfrak{d}_{k,t} = \mathfrak{d}_{k-1,t} - c_{k,t} + 1 \tag{1.115}$$

and

$$\begin{aligned}&\forall\, k\in\{1,2,\dots,L\},\ n\in\{1,2,\dots,l_k\}\colon\\ &x_{k,n} = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k)+\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(k),\mathfrak{d}_{k,1},\mathfrak{d}_{k,2},\dots,\mathfrak{d}_{k,T}}(B_{k,n}\mathbf{I}^{\mathfrak{d}_{k,1},\mathfrak{d}_{k,2},\dots,\mathfrak{d}_{k,T}} + \textstyle\sum_{m=1}^{l_{k-1}} x_{k-1,m} * W_{k,n,m})\end{aligned} \tag{1.116}$$

that

$$(\mathcal{R}^{\mathbf{C}}_a(\Phi))(x_0) = x_L \tag{1.117}$$

and we call $\mathcal{R}^{\mathbf{C}}_a(\Phi)$ the realization function of the feedforward CNN Φ with activation function a (we call $\mathcal{R}^{\mathbf{C}}_a(\Phi)$ the realization of the feedforward CNN Φ with activation a) (cf. Definitions 1.2.1, 1.4.1, 1.4.2, and 1.4.4).

```
import torch
import torch.nn as nn


```

```python
class ConvolutionalANN(nn.Module):
    def __init__(self):
        super().__init__()
        # The convolutional layer defined here takes any tensor of
        # shape (1, n, m) [a single input] or (N, 1, n, m) [a batch
        # of N inputs] where N, n, m are natural numbers satisfying
        # n >= 3 and m >= 3.
        self.conv1 = nn.Conv2d(
            in_channels=1, out_channels=5, kernel_size=(3, 3)
        )
        self.activation1 = nn.ReLU()
        self.conv2 = nn.Conv2d(
            in_channels=5, out_channels=5, kernel_size=(5, 3)
        )

    def forward(self, x0):
        x1 = self.activation1(self.conv1(x0))
        print(x1.shape)
        x2 = self.conv2(x1)
        print(x2.shape)
        return x2


model = ConvolutionalANN()
x0 = torch.rand(1, 20, 20)
# This will print the shapes of the outputs of the two layers of
# the model, in this case:
# torch.Size([5, 18, 18])
# torch.Size([5, 14, 16])
model(x0)
```

Source code 1.18 (code/conv-ann.py): PYTHON code implementing a feedforward CNN in PYTORCH. The implemented model here corresponds to a feedforward CNN $\Phi \in \mathbf{C}$ where $T = 2$, $L = 2$, $l_0 = 1$, $l_1 = 5$, $l_2 = 5$, $(c_{1,1}, c_{1,2}) = (3, 3)$, $(c_{2,1}, c_{2,2}) = (5, 3)$, and $\Phi \in \left(\bigtimes_{k=1}^{L} \left((\mathbb{R}^{c_{k,1} \times c_{k,2} \times \ldots \times c_{k,T}})^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}\right)\right) = ((\mathbb{R}^{3\times 3})^{5\times 1} \times \mathbb{R}^5) \times ((\mathbb{R}^{3\times 5})^{5\times 5} \times \mathbb{R}^5)$. The model, given an input of shape $(1, d_1, d_2)$ with $d_1 \in \mathbb{N} \cap [7, \infty)$, $d_2 \in \mathbb{N} \cap [5, \infty)$, produces an output of shape $(5, d_1 - 6, d_2 - 4)$, (corresponding to the realization function $\mathcal{R}_a^{\mathbf{C}}(\Phi)$ for $a \in C(\mathbb{R}, \mathbb{R})$ having domain $\bigcup_{d_1, d_2 \in \mathbb{N}, d_1 \geq 7, d_2 \geq 5} (\mathbb{R}^{d_1 \times d_2})^1$ and satisfying for all $d_1 \in \mathbb{N} \cap [7, \infty)$, $d_2 \in \mathbb{N} \cap [5, \infty)$, $x_0 \in (\mathbb{R}^{d_1 \times d_2})^1$ that $(\mathcal{R}_a^{\mathbf{C}}(\Phi))(x_0) \in (\mathbb{R}^{d_1 - 6, d_2 - 4})^5$).

Example 1.4.6 (Example for Definition 1.4.5)**.** *Let $T = 2$, $L = 2$, $l_0 = 1$, $l_1 = 2$, $l_2 = 1$,*

$c_{1,1} = 2$, $c_{1,2} = 2$, $c_{2,1} = 1$, $c_{2,2} = 1$ *and let*

$$\Phi \in \left(\bigtimes_{k=1}^{L} \left((\mathbb{R}^{c_{k,1} \times c_{k,2} \times \ldots \times c_{k,T}})^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k} \right) \right) = \left((\mathbb{R}^{2\times 2})^{2\times 1} \times \mathbb{R}^2 \right) \times \left((\mathbb{R}^{1\times 1})^{1\times 2} \times \mathbb{R}^1 \right) \tag{1.118}$$

satisfy

$$\Phi = \left(\left(\left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right), \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right), \begin{pmatrix} \begin{pmatrix} -2 & (2) \end{pmatrix}, (3) \end{pmatrix} \right). \tag{1.119}$$

Then

$$\left(\mathcal{R}_{\mathfrak{r}}^{\mathbf{C}}(\Phi) \right) \left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \right) = \begin{pmatrix} 11 & 15 \\ 23 & 27 \end{pmatrix} \tag{1.120}$$

(cf. Definitions 1.2.4 and 1.4.5).

Proof for Example 1.4.6. Throughout this proof, let $x_0 \in \mathbb{R}^{3\times 3}$, $x_1 = (x_{1,1}, x_{1,2}) \in (\mathbb{R}^{2\times 2})^2$, $x_2 \in \mathbb{R}^{2\times 2}$ with satisfy that

$$x_0 = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \qquad x_{1,1} = \mathfrak{M}_{\mathfrak{r},2,2}\left(\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right), \tag{1.121}$$

$$x_{1,2} = \mathfrak{M}_{\mathfrak{r},2,2}\left((-1)\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right), \tag{1.122}$$

$$\text{and} \qquad x_2 = \mathfrak{M}_{\mathrm{id}_{\mathbb{R}},2,2}\big(3\mathbf{I}^{2,2} + x_{1,1} * \big(-2\big) + x_{1,2} * \big(2\big) \big). \tag{1.123}$$

Note that (1.117), (1.119), (1.121), (1.122), and (1.123) imply that

$$\left(\mathcal{R}_{\mathfrak{r}}^{\mathbf{C}}(\Phi) \right) \left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \right) = \big(\mathcal{R}_{\mathfrak{r}}^{\mathbf{C}}(\Phi) \big)(x_0) = x_2. \tag{1.124}$$

Next observe that (1.121) ensures that

$$\begin{aligned} x_{1,1} &= \mathfrak{M}_{\mathfrak{r},2\times 2}\left(\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right) = \mathfrak{M}_{\mathfrak{r},2\times 2}\left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right) \\ &= \mathfrak{M}_{\mathfrak{r},2\times 2}\left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \end{aligned} \tag{1.125}$$

Furthermore, note that (1.122) establishes that

$$\begin{aligned} x_{1,2} &= \mathfrak{M}_{\mathfrak{r},2\times 2}\left((-1)\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},2\times 2}\left(\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} + \begin{pmatrix} 6 & 8 \\ 12 & 14 \end{pmatrix}\right) \\ &= \mathfrak{M}_{\mathfrak{r},2\times 2}\left(\begin{pmatrix} 5 & 7 \\ 11 & 13 \end{pmatrix}\right) = \begin{pmatrix} 5 & 7 \\ 11 & 13 \end{pmatrix}. \end{aligned} \tag{1.126}$$

Moreover, observe that this, (1.125), and (1.123) demonstrate that

$$\begin{aligned} x_2 &= \mathfrak{M}_{\mathrm{id}_{\mathbb{R}},2\times 2}\left(3\mathbf{I}^{2,2} + x_{1,1} * (-2) + x_{1,2} * (2)\right) \\ &= \mathfrak{M}_{\mathrm{id}_{\mathbb{R}},2\times 2}\left(3\mathbf{I}^{2,2} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} * (-2) + \begin{pmatrix} 5 & 7 \\ 11 & 13 \end{pmatrix} * (2)\right) \\ &= \mathfrak{M}_{\mathrm{id}_{\mathbb{R}},2\times 2}\left(\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} + \begin{pmatrix} -2 & -2 \\ -2 & -2 \end{pmatrix} + \begin{pmatrix} 10 & 14 \\ 22 & 26 \end{pmatrix}\right) \\ &= \mathfrak{M}_{\mathrm{id}_{\mathbb{R}},2\times 2}\left(\begin{pmatrix} 11 & 15 \\ 23 & 27 \end{pmatrix}\right) = \begin{pmatrix} 11 & 15 \\ 23 & 27 \end{pmatrix}. \end{aligned} \tag{1.127}$$

This and (1.124) establish (1.120). The proof for Example 1.4.6 is thus complete. □

```python
import torch
import torch.nn as nn


model = nn.Sequential(
    nn.Conv2d(in_channels=1, out_channels=2, kernel_size=(2, 2)),
    nn.ReLU(),
    nn.Conv2d(in_channels=2, out_channels=1, kernel_size=(1, 1)),
)

with torch.no_grad():
    model[0].weight.set_(
        torch.Tensor([[[[0, 0], [0, 0]]], [[[1, 0], [0, 1]]]])
    )
    model[0].bias.set_(torch.Tensor([1, -1]))
    model[2].weight.set_(torch.Tensor([[[[-2]], [[2]]]]))
    model[2].bias.set_(torch.Tensor([3]))

x0 = torch.Tensor([[[1, 2, 3], [4, 5, 6], [7, 8, 9]]])
print(model(x0))
```

Source code 1.19 (code/conv-ann-ex.py): PYTHON code implementing the feedforward CNN Φ from Example 1.4.6 (see (1.119)) in PYTORCH and verifying (1.120).

Exercise 1.4.1. Let

$$\Phi = \big(((W_{1,n,m})_{(n,m)\in\{1,2,3\}\times\{1\}}, (B_{1,n})_{n\in\{1,2,3\}}), \\ ((W_{2,n,m})_{(n,m)\in\{1\}\times\{1,2,3\}}, (B_{2,n})_{n\in\{1\}})\big) \in ((\mathbb{R}^2)^{3\times 1}\times\mathbb{R}^3)\times((\mathbb{R}^3)^{1\times 3}\times\mathbb{R}^1) \tag{1.128}$$

satisfy

$$W_{1,1,1} = (1,-1),\quad W_{1,2,1} = (2,-2),\quad W_{1,3,1} = (-3,3),\quad (B_{1,n})_{n\in\{1,2,3\}} = (1,2,3), \tag{1.129}$$

$$W_{2,1,1} = (1,-1,1),\quad W_{2,1,2} = (2,-2,2),\quad W_{2,1,3} = (-3,3,-3),\quad \text{and}\quad B_{2,1} = -2 \tag{1.130}$$

and let $v \in \mathbb{R}^9$ satisfy $v = (1,2,3,4,5,4,3,2,1)$. Specify

$$(\mathcal{R}^{\mathbf{C}}_{\mathfrak{r}}(\Phi))(v) \tag{1.131}$$

explicitly and prove that your result is correct (cf. Definitions 1.2.4 and 1.4.5)!

Exercise 1.4.2. Let

$$\Phi = \big(((W_{1,n,m})_{(n,m)\in\{1,2,3\}\times\{1\}}, (B_{1,n})_{n\in\{1,2,3\}}), \\ ((W_{2,n,m})_{(n,m)\in\{1\}\times\{1,2,3\}}, (B_{2,n})_{n\in\{1\}})\big) \in ((\mathbb{R}^3)^{3\times 1}\times\mathbb{R}^3)\times((\mathbb{R}^2)^{1\times 3}\times\mathbb{R}^1) \tag{1.132}$$

satisfy

$$W_{1,1,1} = (1,1,1),\qquad W_{1,2,1} = (2,-2,-2), \tag{1.133}$$

$$W_{1,3,1} = (-3,-3,3),\qquad (B_{1,n})_{n\in\{1,2,3\}} = (3,-2,-1), \tag{1.134}$$

$$W_{2,1,1} = (2,-1),\quad W_{2,1,2} = (-1,2),\quad W_{2,1,3} = (-1,0),\quad \text{and}\quad B_{2,1} = -2 \tag{1.135}$$

and let $v \in \mathbb{R}^9$ satisfy $v = (1,-1,1,-1,1,-1,1,-1,1)$. Specify

$$(\mathcal{R}^{\mathbf{C}}_{\mathfrak{r}}(\Phi))(v) \tag{1.136}$$

explicitly and prove that your result is correct (cf. Definitions 1.2.4 and 1.4.5)!

Exercise 1.4.3. Prove or disprove the following statement: For every $a \in C(\mathbb{R},\mathbb{R})$, $\Phi \in \mathbf{N}$ there exists $\Psi \in \mathbf{C}$ such that for all $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ it holds that $\mathbb{R}^{\mathcal{I}(\Phi)} \subseteq \mathrm{Domain}(\mathcal{R}^{\mathbf{C}}_a(\Psi))$ and

$$(\mathcal{R}^{\mathbf{C}}_a(\Psi))(x) = (\mathcal{R}^{\mathbf{N}}_a(\Phi))(x) \tag{1.137}$$

(cf. Definitions 1.3.1, 1.3.4, 1.4.2, and 1.4.5).

Definition 1.4.7 (Standard scalar products). *We denote by* $\langle \cdot, \cdot \rangle \colon \left[\bigcup_{d \in \mathbb{N}} (\mathbb{R}^d \times \mathbb{R}^d)\right] \to \mathbb{R}$ *the function which satisfies for all* $d \in \mathbb{N}$, $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$ *that*

$$\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i. \tag{1.138}$$

Exercise 1.4.4. For every $d \in \mathbb{N}$ let $\mathbf{e}_1^{(d)}, \mathbf{e}_2^{(d)}, \ldots, \mathbf{e}_d^{(d)} \in \mathbb{R}^d$ satisfy $\mathbf{e}_1^{(d)} = (1, 0, \ldots, 0)$, $\mathbf{e}_2^{(d)} = (0, 1, 0, \ldots, 0)$, $\ldots$, $\mathbf{e}_d^{(d)} = (0, \ldots, 0, 1)$. Prove or disprove the following statement: For all $a \in C(\mathbb{R}, \mathbb{R})$, $\Phi \in \mathbf{N}$, $D \in \mathbb{N}$, $x = ((x_{i,j})_{j \in \{1,2,\ldots,D\}})_{i \in \{1,2,\ldots,\mathcal{I}(\Phi)\}} \in (\mathbb{R}^D)^{\mathcal{I}(\Phi)}$ it holds that

$$(\mathcal{R}_a^{\mathbf{C}}(\Phi))(x) = \left(\left(\langle \mathbf{e}_k^{(\mathcal{O}(\Phi))}, (\mathcal{R}_a^{\mathbf{N}}(\Phi))((x_{i,j})_{i \in \{1,2,\ldots,\mathcal{I}(\Phi)\}}) \rangle\right)_{j \in \{1,2,\ldots,D\}}\right)_{k \in \{1,2,\ldots,\mathcal{O}(\Phi)\}} \tag{1.139}$$

(cf. Definitions 1.3.1, 1.3.4, 1.4.5, and 1.4.7).

1.5 Residual ANNs (ResNets)

In this section we review ResNets. Roughly speaking, plain-vanilla feedforward ANNs can be seen as having a computational structure consisting of sequentially chained layers in which each layer feeds information forward to the next layer (cf., for example, Definitions 1.1.3 and 1.3.4 above). ResNets, in turn, are ANNs involving so-called *skip connections* in their computational structure, which allow information from one layer to be fed not only to the next layer, but also to other layers further down the computational structure. In principle, such skip connections can be employed in combinations with other ANN architecture elements, such as fully-connected layers (cf., for instance, Sections 1.1 and 1.3 above), convolutional layers (cf., for example, Section 1.4 above), and recurrent structures (cf., for instance, Section 1.6 below). However, for simplicity we introduce in this section in all mathematical details feedforward fully-connected ResNets in which the skip connection is a learnable linear map (see Definitions 1.5.1 and 1.5.4 below).

ResNets were introduced in He et al. [198] as an attempt to improve the performance of deep ANNs which typically are much harder to train than shallow ANNs (cf., for example, [30, 159, 342]). The ResNets in He et al. [198] only involve skip connections that are identity mappings without trainable parameters, and are thus a special case of the definition of ResNets provided in this section (see Definitions 1.5.1 and 1.5.4 below). The idea of skip connection (sometimes also called *shortcut connections*) has already been introduced before ResNets and has been used in earlier ANN architecture such as the *highway nets* in Srivastava et al. [398, 399] (cf. also [274, 306, 359, 404, 412]). In addition, we refer to [199, 214, 418, 431, 441] for a few successful ANN architecures building on the ResNets in He et al. [198].

1.5.1 Structured description of fully-connected ResNets

Definition 1.5.1 (Structured description of fully-connected ResNets)**.** *We denote by* $\mathbf{R}$ *the set given by*

$$\mathbf{R} = \bigcup_{L\in\mathbb{N}} \bigcup_{l_0,l_1,\dots,l_L\in\mathbb{N}} \bigcup_{S\subseteq\{(r,k)\in(\mathbb{N}_0)^2\colon r<k\leq L\}} \Big(\big(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k\times l_{k-1}} \times \mathbb{R}^{l_k})\big) \times \Big(\bigtimes_{(r,k)\in S} \mathbb{R}^{l_k\times l_r}\Big)\Big). \tag{1.140}$$

Definition 1.5.2 (Fully-connected ResNets)**.** *We say that* Φ *is a fully-connected ResNet if and only if it holds that*

$$\Phi \in \mathbf{R} \tag{1.141}$$

(cf. Definition 1.5.1).

Lemma 1.5.3 (On an empty set of skip connections)**.** *Let* $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $S \subseteq \{(r,k) \in (\mathbb{N}_0)^2 \colon r < k \leq L\}$. *Then*

$$\#\big(\bigtimes_{(r,k)\in S} \mathbb{R}^{l_k\times l_r}\big) = \begin{cases} 1 & : S = \emptyset \\ \infty & : S \neq \emptyset. \end{cases} \tag{1.142}$$

Proof of Lemma 1.5.3. Throughout this proof, for all sets A and B let $F(A,B)$ be the set of all functions from A to B. Note that

$$\#\big(\bigtimes_{(r,k)\in S} \mathbb{R}^{l_k\times l_r}\big) = \#\big\{f \in F\big(S, \cup_{(r,k)\in S}\mathbb{R}^{l_k\times l_r}\big) \colon (\forall\,(r,k)\in S\colon f(r,k) \in \mathbb{R}^{l_k\times l_r})\big\}. \tag{1.143}$$

This and the fact that for all sets B it holds that $\#(F(\emptyset,B)) = 1$ show that

$$\#\big(\bigtimes_{(r,k)\in\emptyset} \mathbb{R}^{l_k\times l_r}\big) = \#(F(\emptyset,\emptyset)) = 1. \tag{1.144}$$

Next note that (1.143) establishes that for all $(R,K) \in S$ it holds that

$$\#\big(\bigtimes_{(r,k)\in S} \mathbb{R}^{l_k\times l_r}\big) \geq \#\big(F\big(\{(R,K)\}, \mathbb{R}^{l_K\times l_R}\big)\big) = \infty. \tag{1.145}$$

Combining this and (1.144) establishes (1.142). The proof of Lemma 1.5.3 is thus complete. □

1.5.2 Realizations of fully-connected ResNets

Definition 1.5.4 (Realizations associated to fully-connected ResNets)**.** *Let $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$, $S \subseteq \{(r,k) \in (\mathbb{N}_0)^2 \colon r < k \leq L\}$, $\Phi = ((W_k, B_k)_{k \in \{1,2,\ldots,L\}}, (V_{r,k})_{(r,k)\in S}) \in \big(\big(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\big) \times \big(\bigtimes_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}\big)\big) \subseteq \mathbf{R}$ and let $a \colon \mathbb{R} \to \mathbb{R}$ be a function. Then we denote by*

$$\mathcal{R}_a^{\mathbf{R}}(\Phi) \colon \mathbb{R}^{l_0} \to \mathbb{R}^{l_L} \tag{1.146}$$

the function which satisfies for all $x_0 \in \mathbb{R}^{l_0}, x_1 \in \mathbb{R}^{l_1}, \ldots, x_L \in \mathbb{R}^{l_L}$ with

$$\forall\, k \in \{1, 2, \ldots, L\} \colon x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k) + \mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(k), l_k}(W_k x_{k-1} + B_k + \textstyle\sum_{r\in\mathbb{N}_0,(r,k)\in S} V_{r,k} x_r) \tag{1.147}$$

that

$$(\mathcal{R}_a^{\mathbf{R}}(\Phi))(x_0) = x_L \tag{1.148}$$

and we call $\mathcal{R}_a^{\mathbf{R}}(\Phi)$ the realization function of the fully-connected ResNet Φ with activation function a (we call $\mathcal{R}_a^{\mathbf{R}}(\Phi)$ the realization of the fully-connected ResNet Φ with activation a) (cf. Definitions 1.2.1 and 1.5.1).

Definition 1.5.5 (Identity matrices)**.** *Let $d \in \mathbb{N}$. Then we denote by $\mathrm{I}_d \in \mathbb{R}^{d\times d}$ the identity matrix in $\mathbb{R}^{d\times d}$.*

```python
import torch
import torch.nn as nn

class ResidualANN(nn.Module):
    def __init__(self):
        super().__init__()
        self.affine1 = nn.Linear(3, 10)
        self.activation1 = nn.ReLU()
        self.affine2 = nn.Linear(10, 20)
        self.activation2 = nn.ReLU()
        self.affine3 = nn.Linear(20, 10)
        self.activation3 = nn.ReLU()
        self.affine4 = nn.Linear(10, 1)

    def forward(self, x0):
        x1 = self.activation1(self.affine1(x0))
        x2 = self.activation2(self.affine2(x1))
        x3 = self.activation3(x1 + self.affine3(x2))
        x4 = self.affine4(x3)
        return x4
```

Source code 1.20 (code/res-ann.py): PYTHON code implementing a fully-connected ResNet in PYTORCH. The implemented model here corresponds to a fully-connected ResNet (Φ, V) where $l_0 = 3$, $l_1 = 10$, $l_2 = 20$, $l_3 = 10$, $l_4 = 1$, $\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3), (W_4, B_4)) \in \left(\bigtimes_{k=1}^{4} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right)$, $S = \{(1,3)\}$, $V = (V_{r,k})_{(r,k)\in S} \in \left(\bigtimes_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}\right)$, and $V_{1,3} = \mathrm{I}_{10}$ (cf. Definition 1.5.5).

Example 1.5.6 (Example for Definition 1.5.2)**.** *Let $l_0 = 1$, $l_1 = 1$, $l_2 = 2$, $l_3 = 2$, $l_4 = 1$, $S = \{(0,4)\}$, let*

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3), (W_4, B_4)) \in \left(\bigtimes_{k=1}^{4} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right) \tag{1.149}$$

satisfy

$$W_1 = \begin{pmatrix} 1 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} 0 \end{pmatrix}, \qquad W_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{1.150}$$

$$W_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad W_4 = \begin{pmatrix} 2 & 2 \end{pmatrix}, \qquad \text{and} \qquad B_4 = \begin{pmatrix} 1 \end{pmatrix}, \tag{1.151}$$

and let $V = (V_{r,k})_{(r,k)\in S} \in \bigtimes_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}$ satisfy

$$V_{0,4} = \begin{pmatrix} -1 \end{pmatrix}. \tag{1.152}$$

Then

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(5) = 28 \tag{1.153}$$

(cf. Definitions 1.2.4 and 1.5.4).

Proof for Example 1.5.6. Throughout this proof, let $x_0 \in \mathbb{R}^1$, $x_1 \in \mathbb{R}^1$, $x_2 \in \mathbb{R}^2$, $x_3 \in \mathbb{R}^2$, $x_4 \in \mathbb{R}^1$ satisfy for all $k \in \{1,2,3,4\}$ that $x_0 = 5$ and

$$x_k = \mathfrak{M}_{\mathfrak{r}\mathbb{1}_{(0,4)}(k)+\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{4\}}(k),l_k}(W_k x_{k-1} + B_k + \textstyle\sum_{r\in\mathbb{N}_0,(r,k)\in S} V_{r,k} x_r). \tag{1.154}$$

Observe that (1.154) shows that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(5) = x_4. \tag{1.155}$$

Next note that (1.154) ensures that

$$x_1 = \mathfrak{M}_{\mathfrak{r},1}(W_1 x_0 + B_1) = \mathfrak{M}_{\mathfrak{r},1}(5), \tag{1.156}$$

$$x_2 = \mathfrak{M}_{\mathfrak{r},2}(W_2 x_1 + B_2) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix}1\\2\end{pmatrix}(5) + \begin{pmatrix}0\\1\end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix}5\\11\end{pmatrix}\right) = \begin{pmatrix}5\\11\end{pmatrix}, \tag{1.157}$$

$$x_3 = \mathfrak{M}_{\mathfrak{r},2}(W_3 x_2 + B_3) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\begin{pmatrix}5\\11\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix}5\\11\end{pmatrix}\right) = \begin{pmatrix}5\\11\end{pmatrix}, \tag{1.158}$$

$$\begin{aligned} \text{and} \qquad x_4 &= \mathfrak{M}_{\mathfrak{r},1}(W_4 x_3 + B_4 + V_{0,4} x_0) \\ &= \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix}2 & 2\end{pmatrix}\begin{pmatrix}5\\11\end{pmatrix} + (1) + (-1)(5)\right) = \mathfrak{M}_{\mathfrak{r},1}(28) = 28. \end{aligned} \tag{1.159}$$

This and (1.155) establish (1.153). The proof for Example 1.5.6 is thus complete. □

Exercise 1.5.1. Let $l_0 = 1$, $l_1 = 2$, $l_2 = 3$, $l_3 = 1$, $S = \{(0,3),(1,3)\}$, let

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3)) \in \left(\bigtimes_{k=1}^{3} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\right) \tag{1.160}$$

satisfy

$$W_1 = \begin{pmatrix}1\\2\end{pmatrix}, \qquad B_1 = \begin{pmatrix}3\\4\end{pmatrix}, \qquad W_2 = \begin{pmatrix}-1 & 2\\3 & -4\\-5 & 6\end{pmatrix}, \qquad B_2 = \begin{pmatrix}0\\0\\0\end{pmatrix}, \tag{1.161}$$

$$W_3 = \begin{pmatrix}-1 & 1 & -1\end{pmatrix}, \qquad \text{and} \qquad B_3 = \begin{pmatrix}-4\end{pmatrix}, \tag{1.162}$$

and let $V = (V_{r,k})_{(r,k)\in S} \in \bigtimes_{(r,k)\in S} \mathbb{R}^{l_k \times l_r}$ satisfy

$$V_{0,3} = \begin{pmatrix}1\end{pmatrix} \qquad \text{and} \qquad V_{1,3} = \begin{pmatrix}3 & -2\end{pmatrix}. \tag{1.163}$$

Prove or disprove the following statement: It holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(-1) = 0 \tag{1.164}$$

(cf. Definitions 1.2.4 and 1.5.4).

1.6 Recurrent ANNs (RNNs)

In this section we review RNNs, a type of ANNs designed to take sequences of data points as inputs. Roughly speaking, unlike in feedforward ANNs where an input is processed by a successive application of series of *different* parametric functions (cf. Definitions 1.1.3, 1.3.4, 1.4.5, and 1.5.4 above), in RNNs an input sequence is processed by a repeated application of the *same* parametric function whereby after the first application, each subsequent application of the parametric function takes as input a new element of the input sequence and a partial output from the previous application of the parametric function. The output of an RNN is then given by a sequence of partial outputs coming from the

repeated applications of the parametric function (see Definition 1.6.2 below for a precise description of RNNs and cf., for instance, [4, Section 12.7], [61, Chapter 17] [64, Chapter 5], and [170, Chapter 10] for other introductions to RNNs).

The repeatedly applied parametric function in an RNN is typically called an *RNN node* and any RNN architecture is determined by specifying the architecture of the corresponding RNN node. We review a simple variant of such RNN nodes and the corresponding RNNs in Section 1.6.2 in detail and we briefly address one of the most commonly used RNN nodes, the so-called *long short-term memory* (LSTM) node, in Section 1.6.3.

There is a wide range of application areas where sequential data are considered and RNN based deep learning methods are being employed and developed. Examples of such applications areas are NLP including language translation (cf., for example, [11, 78, 79, 402] and the references therein), language generation (cf., for instance, [52, 175, 248, 354] and the references therein), and speech recognition (cf., for example, [6, 83, 176, 178, 374] and the references therein), time series prediction analysis including stock market prediction (cf., for instance, [135, 138, 386, 390] and the references therein) and weather prediction (cf., for example, [366, 389, 421] and the references therein) and video analysis (cf., for instance, [113, 245, 321, 415] and the references therein).

1.6.1 Description of RNNs

Definition 1.6.1 (Function unrolling)**.** *Let X, Y, I be sets, let $f\colon X \times I \to Y \times I$ be a function, and let $T \in \mathbb{N}$, $\mathbb{I} \in I$. Then we denote by $\mathfrak{R}_{f,T,\mathbb{I}}\colon X^T \to Y^T$ the function which satisfies for all $x_1, x_2, \ldots, x_T \in X$, $y_1, y_2, \ldots, y_T \in Y$, $i_0, i_1, \ldots, i_T \in I$ with $i_0 = \mathbb{I}$ and $\forall\, t \in \{1, 2, \ldots, T\}\colon (y_t, i_t) = f(x_t, i_{t-1})$ that*

$$\mathfrak{R}_{f,T,\mathbb{I}}(x_1, x_2, \ldots, x_T) = (y_1, y_2, \ldots, y_T) \tag{1.165}$$

and we call $\mathfrak{R}_{f,T,\mathrm{i}}$ the T-times unrolled function f with initial information $\mathbb{I}$.

Definition 1.6.2 (Description of RNNs)**.** *Let X, Y, I be sets, let $\mathfrak{d}, T \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{I} \in I$, and let $\mathfrak{N} = (\mathfrak{N}_\vartheta)_{\vartheta \in \mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \times X \times I \to Y \times I$ be a function. Then we call R the realization function of the T-step unrolled RNN with RNN node $\mathfrak{N}$, parameter vector θ, and initial information $\mathbb{I}$ (we call R the realization of the T-step unrolled RNN with RNN node $\mathfrak{N}$, parameter vector θ, and initial information $\mathbb{I}$) if and only if it holds that*

$$R = \mathfrak{R}_{\mathfrak{N}_\theta,T,\mathbb{I}} \tag{1.166}$$

(cf. Definition 1.6.1).

1.6.2 Vectorized description of simple fully-connected RNNs

Definition 1.6.3 (Vectorized description of simple fully-connected RNN nodes)**.** *Let $\mathfrak{r}, \mathfrak{y}, \mathfrak{i} \in \mathbb{N}$, $\theta \in \mathbb{R}^{(\mathfrak{r}+\mathfrak{i}+1)\mathfrak{i}+(\mathfrak{i}+1)\mathfrak{y}}$ and let $\Psi_1 \colon \mathbb{R}^{\mathfrak{i}} \to \mathbb{R}^{\mathfrak{i}}$ and $\Psi_2 \colon \mathbb{R}^{\mathfrak{y}} \to \mathbb{R}^{\mathfrak{y}}$ be functions. Then we call r the realization function of the simple fully-connected RNN node with parameter vector θ and activation functions Ψ_1 and Ψ_2 (we call r the realization of the simple fully-connected RNN node with parameter vector θ and activations Ψ_1 and Ψ_2) if and only if it holds that $r \colon \mathbb{R}^{\mathfrak{r}} \times \mathbb{R}^{\mathfrak{i}} \to \mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ is the function from $\mathbb{R}^{\mathfrak{r}} \times \mathbb{R}^{\mathfrak{i}}$ to $\mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ which satisfies for all $x \in \mathbb{R}^{\mathfrak{r}}$, $i \in \mathbb{R}^{\mathfrak{i}}$ that*

$$r(x, i) = \Big(\big(\mathcal{N}^{\theta, \mathfrak{r}+\mathfrak{i}}_{\Psi_1, \Psi_2} \big)(x, i), \big(\mathcal{N}^{\theta, \mathfrak{r}+\mathfrak{i}}_{\Psi_1} \big)(x, i) \Big) \tag{1.167}$$

(cf. Definition 1.1.3).

Definition 1.6.4 (Vectorized description of simple fully-connected RNNs)**.** *Let $\mathfrak{r}, \mathfrak{y}, \mathfrak{i}, T \in \mathbb{N}$, $\theta \in \mathbb{R}^{(\mathfrak{r}+\mathfrak{i}+1)\mathfrak{i}+(\mathfrak{i}+1)\mathfrak{y}}$, $\mathbb{I} \in \mathbb{R}^{\mathfrak{i}}$ and let $\Psi_1 \colon \mathbb{R}^{\mathfrak{i}} \to \mathbb{R}^{\mathfrak{i}}$ and $\Psi_2 \colon \mathbb{R}^{\mathfrak{y}} \to \mathbb{R}^{\mathfrak{y}}$ be functions. Then we call R the realization function of the T-step unrolled simple fully-connected RNN with parameter vector θ, activation functions Ψ_1 and Ψ_2, and initial information $\mathbb{I}$ (we call R the realization of the T-step unrolled simple fully-connected RNN with parameter vector θ, activations Ψ_1 and Ψ_2, and initial information $\mathbb{I}$) if and only if there exists $r \colon \mathbb{R}^{\mathfrak{r}} \times \mathbb{R}^{\mathfrak{i}} \to \mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ such that*

(i) it holds that r is the realization of the simple fully-connected RNN node with parameter vector θ and activations Ψ_1 and Ψ_2 and

(ii) it holds that

$$R = \mathfrak{R}_{r, T, \mathbb{I}} \tag{1.168}$$

(cf. Definitions 1.6.1 and 1.6.3).

Lemma 1.6.5. *Let $\mathfrak{r}, \mathfrak{y}, \mathfrak{i}, \mathfrak{d}, T \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{I} \in \mathbb{R}^{\mathfrak{i}}$ satisfy $\mathfrak{d} = (\mathfrak{r}+\mathfrak{i}+1)\mathfrak{i} + (\mathfrak{i}+1)\mathfrak{y}$, let $\Psi_1 \colon \mathbb{R}^{\mathfrak{i}} \to \mathbb{R}^{\mathfrak{i}}$ and $\Psi_2 \colon \mathbb{R}^{\mathfrak{y}} \to \mathbb{R}^{\mathfrak{y}}$ be functions, and let $\mathfrak{N} = (\mathfrak{N}_\vartheta)_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{r}} \times \mathbb{R}^{\mathfrak{i}} \to \mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ satisfy for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathfrak{N}_\vartheta$ is the realization of the simple fully-connected RNN node with parameter vector ϑ and activations Ψ_1 and Ψ_2 (cf. Definition 1.6.3). Then the following three statements are equivalent:*

(i) It holds that R is the realization of the T-step unrolled simple fully-connected RNN with parameter vector θ, activations Ψ_1 and Ψ_2, and initial information $\mathbb{I}$ (cf.

Definition 1.6.4).

(ii) It holds that R is the realization of the T-step unrolled RNN with RNN node $\mathfrak{N}$, parameter vector θ, and initial information $\mathbb{I}$ (cf. Definition 1.6.2).

(iii) It holds that

$$R = \mathfrak{R}_{\mathfrak{N}_\theta, T, \mathbb{I}} \tag{1.169}$$

(cf. Definition 1.6.1).

Proof of Lemma 1.6.5. Observe that (1.166), (1.168), and (1.169) prove that ((i) $\leftrightarrow$ (ii) $\leftrightarrow$ (iii)). The proof of Lemma 1.6.5 is thus complete. □

Exercise 1.6.1. For every $T \in \mathbb{N}$, $\alpha \in (0,1)$ let $R_{T,\alpha}$ be the realization of the T-step unrolled simple fully-connected RNN with parameter vector $(1, 0, 0, \alpha, 0, 1-\alpha, 0, 0, -1, 1, 0)$, activations $\mathfrak{M}_{\mathfrak{r},2}$ and $\mathrm{id}_{\mathbb{R}}$, and initial information $(0, 0)$ (cf. Definitions 1.2.1, 1.2.4, and 1.6.4). For every $T \in \mathbb{N}$, $\alpha \in (0,1)$ specify $R_{T,\alpha}(1, 1, \ldots, 1)$ explicitly and prove that your result is correct!

1.6.3 Long short-term memory (LSTM) RNNs

In this section we briefly discuss a very popular type of RNN nodes called *LSTM nodes* and the corresponding RNNs called *LSTM networks* which were introduced in Hochreiter & Schmidhuber [209]. Loosely speaking, LSTM nodes were invented to attempt to the tackle the issue that most RNNs based on simple RNN nodes, such as the simple fully-connected RNN nodes in Section 1.6.2 above, struggle to learn to understand long-term dependencies in sequences of data (cf., for example, [30, 342]). Roughly speaking, an RNN processes an input sequence by repeatedly applying an RNN node to a tuple consisting of a new element of the input sequence and a partial output of the previous application of the RNN node (see Definition 1.6.2 above for a precise description of RNNs). Therefore, the only information on previously processed elements of the input sequence that any application of an RNN node has access to, is the information encoded in the output produced by the last application of the RNN node. For this reason, RNNs can be seen as only having a *short-term memory*. The LSTM architecture, however is designed with the aim to facilitate the transmission of long-term information within this short-term memory. LSTM networks can thus be seen as having a sort of *long short-term memory*.

For a precise definition of LSTM networks we refer to the original article Hochreiter & Schmidhuber [209] and, for instance, to the excellent explanations in [138, 175, 333]. For a few selected references on LSTM networks in the literature we refer, for example, to [11, 79, 138, 153, 154, 175, 177–180, 301, 344, 374, 381, 402, 439] and the references therein.

1.7 Further types of ANNs

In this section we present a selection of references and some rough comments on a couple of further popular types of ANNs in the literature which were not discussed in the previous sections of this chapter above.

1.7.1 ANNs with encoder-decoder architectures: autoencoders

In this section we discuss the idea of autoencoders which are based on encoder-decoder ANN architectures. Roughly speaking, the goal of autoencoders is to learn a simplified representation of data points and a way to closely reconstruct the original data points from the simplified representation. The simplified representation of data points is usually called the *encoding* and is obtained by applying an *encoder ANN* to the data points. The approximate reconstruction of the original data points from the encoded representations is, in turn, called the *decoding* and is obtained by applying a *decoder ANN* to the encoded representations. The composition of the encoder ANN with the decoder ANN is called the *autoencoder*. In the simplest situations the encoder ANN and decoder ANN are trained to perform their respective desired functions by training the full autoencoder to be as close to the identity mapping on the data points as possible.

A large number of different architectures and training procedures for autoencoders have been proposed in the literature. In the following we list a selection of a few popular ideas from the scientific literature.

- We refer, for instance, to [50, 206, 208, 263, 370] for foundational references introducing and refining the idea of autoencoders,
- we refer, for example, to [416, 417, 430] for so-called *denoising autoencoders* which add random pertubation to the input data in the training of autoencoders,
- we refer, for instance, to [52, 112, 256] for so-called *variational autoencoders* which use techniques from bayesian statistics in the training of autoencoders,
- we refer, for example, [307, 363] for autoencoders involving convolutions, and
- we refer, for instance, [123, 305] for *adversarial autoencoders* which combine the principles of autoencoders with the paradigm of generative adversarial networks (see Goodfellow et al. [171]).

1.7.2 Transformers and the attention mechanism

In Section 1.6 we reviewed RNNs which are a type of ANNs designed to take sequences of data points as inputs. Very roughly speaking, RNNs process a sequence of data points by sequentially processing one data point of the sequence after the other and thereby

constantly updating an information state encoding previously processed information (see Section 1.6.1 above for a precise description of RNNs). When processing a data point of the sequence, any information coming from earlier data points is thus only available to the RNN through the information state passed on from the previous processing step of the RNN. Consequently, it can be hard for RNNs to learn to understand long-term dependencies in the input sequence. In Section 1.6.3 above, we briefly discussed the LSTM architecture for RNNs which is an architecture for RNNs aimed at giving such RNNs the capacity to indeed learn to understand such long-term dependencies.

Another approach in the literature to design ANN architectures which process sequential data and are capable to efficiently learn to understand long-term dependencies in data sequences is called the *attention mechanism*. Very roughly speaking, in the context of sequences of the data, the attention mechanism aims to give ANNs the capacity to "*pay attention*" to selected parts of the entire input sequence when they are processing a data point of the sequence. The idea for using attention mechanisms in ANNs was first introduced in Bahdanau et al. [11] in the context of RNNs trained for machine translation. In this context the proposed ANN architecture still processes the input sequence sequentially, however past information is not only available through the information state from the previous processing step, but also through the attention mechanism, which can directly extract information from data points far away from the data point being processed.

Likely the most famous ANNs based on the attention mechanism do however not involve any recurrent elements and have been named *Transfomer ANNs* by the authors of the seminal paper Vaswani et al. [411] called "Attention is all you need". Roughly speaking, Transfomer ANNs are designed to process sequences of data by considering the entire input sequence at once and relying only on the attention mechanism to understand dependencies between the data points in the sequence. Transfomer ANNs are the basis for many recently very successful *large language models* (LLMs), such as, *generative pre-trained transformers* (GPTs) in [55, 334, 355, 356] which are the models behind the famous *ChatGPT* application, *Bidirectional Encoder Representations from Transformers* (BERT) models in Devlin et al. [109], and many others (cf., for example, [93, 277, 357, 432, 436] and the references therein).

Beyond the NLP applications for which Transformers and attention mechanisms have been introduced, similar ideas have been employed in several other areas, such as, computer vision (cf., for instance, [114, 250, 289, 418]), protein structure prediction (cf., for example, [242]), multimodal learning (cf., for instance, [295]), and long sequence time-series forecasting (cf., for example, [455]). Moreover, we refer, for instance, to [83, 301], [163, Chapter 17], and [170, Section 12.4.5.1] for explorations and explanations of the attention mechanism in the literature.

1.7.3 Graph neural networks (GNNs)

All ANNs reviewed in the previous sections of this book are designed to take real-valued vectors or sequences of real-valued vectors as inputs. However, there are several learning problems based on data, such as social network data or molecular data, that are not optimally represented by real-valued vectors but are better represented by graphs (see, for example, West [425] for an introduction on graphs). As a consequence, many ANN architectures which can process graphs as inputs, so-called *graph neural networks* (GNNs), have been introduced in the literature.

- We refer, for instance, to [376, 429, 453, 456] for overview articles on GNNs,
- we refer, for example, to [172, 380] for foundational articles for GNNs,
- we refer, for instance, to [413, 440] for applications of attention mechanisms (cf. Section 1.7.2 above) to GNNs,
- we refer, for example, to [56, 97, 426, 438] for GNNs involving convolutions on graphs, and
- we refer, for instance, to [16, 157, 375, 382, 428] for applications of GNNs to problems from the natural sciences.

1.7.4 Neural operators

In this section we review a few popular ANN-type architectures employed in *operator learning*. Roughly speaking, in operator learning one is not interested in learning a map between finite-dimensional euclidean spaces, but in learning a map from a space of functions to a space of functions. Such a map between (typically infinite-dimensional) vector spaces is usually called an *operator*. An example of such a map is the solution operator of an evolutionary PDE which maps the initial condition of the PDE to the corresponding terminal value of the PDE. To approximate/learn operators it is necessary to develop parametrized families of operators, objects which we refer to as *neural operators*. Many different architectures for such neural operators have been proposed in the literature, some of which we now list in the next paragraphs.

One of the most successful neural operator architectures are so-called *Fourier neural operators* (FNOs) introduced in Li et al. [282] (cf. also Kovachki et al. [262]). Very roughly speaking, FNOs are parametric maps on function spaces, which involve transformations on function values as well as on Fourier coefficients. FNOs have been derived based on the neural operators introduced in Li et al. [281, 283] which are based on integral transformations with parametric integration kernels. We refer, for example, to [54, 261, 280, 424] and the references therein for extensions and theoretical results on FNOs.

A simple and successful architecture for neural operators, which is based on a universal approximation theorem for neural operators, are the *deep operator networks* (deepONets) introduced in Lu et al. [296]. Roughly speaking, a deepONet consists of two ANNs that take as input the evaluation point of the output space and input function values at predetermined "sensor" points respectively, and that are joined together by a scalar product to produce the output of the deepONet. We refer, for instance, to [120, 173, 259, 271, 287, 310, 349, 406, 420, 427, 446] for extensions and theoretical results on deepONets. For a comparison between deepONets and FNOs we refer, for example, to Lu et al. [297].

A further natural approach is to employ CNNs (see Section 1.4) to develop neural operator architectures. We refer, for instance, to [192, 200, 254, 364, 457] for such CNN-based neural operators. Finally, we refer, for example, to [68, 96, 100, 140, 141, 237, 284, 288, 314, 358, 383, 433] for further neural operator architectures and theoretical results for neural operators.

Chapter 2

ANN calculus

In this chapter we review certain operations that can be performed on the set of fully-connected feedforward 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 fully-connected feedforward ANNsThe operations and the calculus for fully-connected feedforward ANNs presented in this chapter will be used in Chapters 3 and 4 to establish certain ANN approximation results.

In the literature such operations on ANNs and such kind of calculus on ANNs has been used in many research articles such as [133, 165, 186, 187, 191, 238, 335, 343, 347] and the references therein. The specific presentation of this chapter is based on Grohs et al. [186, 187].

2.1 Compositions of fully-connected feedforward ANNs

2.1.1 Compositions of fully-connected feedforward ANNs

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

$$(\cdot) \bullet (\cdot) \colon \{(\Phi, \Psi) \in \mathbf{N} \times \mathbf{N} \colon \mathcal{I}(\Phi) = \mathcal{O}(\Psi)\} \to \mathbf{N} \tag{2.1}$$

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

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

(cf. Definition 1.3.1).

2.1.2 Elementary properties of compositions of fully-connected feedforward ANNs

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

(i) it holds that

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

(ii) it holds that

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

(iii) it holds that

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

(iv) it holds that

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

and

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

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

(cf. Definitions 1.3.4 and 2.1.1).

Proof of Proposition 2.1.2. Throughout this proof, let $L = \mathcal{L}(\Phi \bullet \Psi)$ and for every $a \in C(\mathbb{R}, \mathbb{R})$ let

$$\begin{aligned} X_a = \big\{ x = (x_0, x_1, \dots, x_L) \in \mathbb{R}^{\mathbb{D}_0(\Phi \bullet \Psi)} \times \mathbb{R}^{\mathbb{D}_1(\Phi \bullet \Psi)} \times \cdots \times \mathbb{R}^{\mathbb{D}_L(\Phi \bullet \Psi)} \colon \\ \big(\forall\, k \in \{1, 2, \dots, L\} \colon x_k = \mathfrak{M}_{a \mathbb{1}_{(0,L)}(k) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), \mathbb{D}_k(\Phi \bullet \Psi)}(\mathcal{W}_{k, \Phi \bullet \Psi} x_{k-1} + \mathcal{B}_{k, \Phi \bullet \Psi})\big)\big\}. \end{aligned} \tag{2.8}$$

Note 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). Observe that item (iii) in Lemma 1.3.3 and (2.2) show that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$\mathcal{W}_{k, \Phi \bullet \Psi} \in \begin{cases} \mathbb{R}^{\mathbb{D}_k(\Psi) \times \mathbb{D}_{k-1}(\Psi)} & : k < \mathcal{L}(\Psi) \\ \mathbb{R}^{\mathbb{D}_1(\Phi) \times \mathbb{D}_{\mathcal{L}(\Psi)-1}(\Psi)} & : k = \mathcal{L}(\Psi) \\ \mathbb{R}^{\mathbb{D}_{k-\mathcal{L}(\Psi)+1}(\Phi) \times \mathbb{D}_{k-\mathcal{L}(\Psi)}(\Phi)} & : k > \mathcal{L}(\Psi). \end{cases} \tag{2.9}$$

This, item (iii) in Lemma 1.3.3, and the fact that $\mathcal{H}(\Psi) = \mathcal{L}(\Psi) - 1$ ensure that for all $k \in \{0, 1, \ldots, L\}$ it holds that

$$\mathbb{D}_k(\Phi \bullet \Psi) = \begin{cases} \mathbb{D}_k(\Psi) & : k \leq \mathcal{H}(\Psi) \\ \mathbb{D}_{k-\mathcal{L}(\Psi)+1}(\Phi) & : k > \mathcal{H}(\Psi). \end{cases} \tag{2.10}$$

This establishes item (i). Note that (2.10) implies that

$$\begin{aligned} \mathcal{P}(\Phi \bullet \Psi) &= \sum_{j=1}^{L} \mathbb{D}_j(\Phi \bullet \Psi)(\mathbb{D}_{j-1}(\Phi \bullet \Psi) + 1) \\ &= \left[\sum_{j=1}^{\mathcal{H}(\Psi)} \mathbb{D}_j(\Psi)(\mathbb{D}_{j-1}(\Psi) + 1)\right] + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi) + 1) \\ &\quad + \left[\sum_{j=\mathcal{L}(\Psi)+1}^{L} \mathbb{D}_{j-\mathcal{L}(\Psi)+1}(\Phi)(\mathbb{D}_{j-\mathcal{L}(\Psi)}(\Phi) + 1)\right] \\ &= \left[\sum_{j=1}^{\mathcal{L}(\Psi)-1} \mathbb{D}_j(\Psi)(\mathbb{D}_{j-1}(\Psi) + 1)\right] + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi) + 1) \\ &\quad + \left[\sum_{j=2}^{\mathcal{L}(\Phi)} \mathbb{D}_j(\Phi)(\mathbb{D}_{j-1}(\Phi) + 1)\right] \\ &= \big[\mathcal{P}(\Psi) - \mathbb{D}_{\mathcal{L}(\Psi)}(\Psi)(\mathbb{D}_{\mathcal{L}(\Psi)-1}(\Psi) + 1)\big] + \mathbb{D}_1(\Phi)(\mathbb{D}_{\mathcal{H}(\Psi)}(\Psi) + 1) \\ &\quad + \big[\mathcal{P}(\Phi) - \mathbb{D}_1(\Phi)(\mathbb{D}_0(\Phi) + 1)\big]. \end{aligned} \tag{2.11}$$

This proves item (iv). Observe that (2.10) and item (ii) in Lemma 1.3.3 ensure that

$$\begin{aligned} &\mathcal{I}(\Phi \bullet \Psi) = \mathbb{D}_0(\Phi \bullet \Psi) = \mathbb{D}_0(\Psi) = \mathcal{I}(\Psi) \\ \text{and} \quad &\mathcal{O}(\Phi \bullet \Psi) = \mathbb{D}_{\mathcal{L}(\Phi \bullet \Psi)}(\Phi \bullet \Psi) = \mathbb{D}_{\mathcal{L}(\Phi \bullet \Psi) - \mathcal{L}(\Psi) + 1}(\Phi) = \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi) = \mathcal{O}(\Phi). \end{aligned} \tag{2.12}$$

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

$$\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \Psi) \in C(\mathbb{R}^{\mathcal{I}(\Phi \bullet \Psi)}, \mathbb{R}^{\mathcal{O}(\Phi \bullet \Psi)}) = C(\mathbb{R}^{\mathcal{I}(\Psi)}, \mathbb{R}^{\mathcal{O}(\Phi)}). \tag{2.13}$$

Next note that (2.2) implies that for all $k \in \mathbb{N} \cap (1, \mathcal{L}(\Phi) + 1)$ it holds that

$$(\mathcal{W}_{\mathcal{L}(\Psi)+k-1, \Phi \bullet \Psi}, \mathcal{B}_{\mathcal{L}(\Psi)+k-1, \Phi \bullet \Psi}) = (\mathcal{W}_{k,\Phi}, \mathcal{B}_{k,\Phi}). \tag{2.14}$$

This and (2.10) ensure that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x = (x_0, x_1, \ldots, x_L) \in X_a$, $k \in \mathbb{N} \cap (1, \mathcal{L}(\Phi) + 1)$ it holds that

$$\begin{aligned} x_{\mathcal{L}(\Psi)+k-1} &= \mathfrak{M}_{a\mathbb{1}_{(0,L)}(\mathcal{L}(\Psi)+k-1) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(\mathcal{L}(\Psi)+k-1), \mathbb{D}_k(\Phi)}(\mathcal{W}_{k,\Phi} x_{\mathcal{L}(\Psi)+k-2} + \mathcal{B}_{k,\Phi}) \\ &= \mathfrak{M}_{a\mathbb{1}_{(0,\mathcal{L}(\Phi))}(k) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{\mathcal{L}(\Phi)\}}(k), \mathbb{D}_k(\Phi)}(\mathcal{W}_{k,\Phi} x_{\mathcal{L}(\Psi)+k-2} + \mathcal{B}_{k,\Phi}). \end{aligned} \tag{2.15}$$

Furthermore, observe that (2.2) and (2.10) show that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x = (x_0, x_1, \ldots, x_L) \in X_a$ it holds that

$$\begin{aligned} x_{\mathcal{L}(\Psi)} &= \mathfrak{M}_{a\mathbb{1}_{(0,L)}(\mathcal{L}(\Psi))+\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(\mathcal{L}(\Psi)),\mathbb{D}_{\mathcal{L}(\Psi)}(\Phi\bullet\Psi)}(\mathcal{W}_{\mathcal{L}(\Psi),\Phi\bullet\Psi}x_{\mathcal{L}(\Psi)-1} + \mathcal{B}_{\mathcal{L}(\Psi),\Phi\bullet\Psi}) \\ &= \mathfrak{M}_{a\mathbb{1}_{(0,\mathcal{L}(\Phi))}(1)+\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{\mathcal{L}(\Phi)\}}(1),\mathbb{D}_1(\Phi)}(\mathcal{W}_{1,\Phi}\mathcal{W}_{\mathcal{L}(\Psi),\Psi}x_{\mathcal{L}(\Psi)-1} + \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Psi),\Psi} + \mathcal{B}_{1,\Phi}) \\ &= \mathfrak{M}_{a\mathbb{1}_{(0,\mathcal{L}(\Phi))}(1)+\mathrm{id}_{\mathbb{R}}\mathbb{1}_{\{\mathcal{L}(\Phi)\}}(1),\mathbb{D}_1(\Phi)}(\mathcal{W}_{1,\Phi}(\mathcal{W}_{\mathcal{L}(\Psi),\Psi}x_{\mathcal{L}(\Psi)-1} + \mathcal{B}_{\mathcal{L}(\Psi),\Psi}) + \mathcal{B}_{1,\Phi}). \end{aligned} \tag{2.16}$$

Combining this and (2.15) proves that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x = (x_0, x_1, \ldots, x_L) \in X_a$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_a(\Phi))(\mathcal{W}_{\mathcal{L}(\Psi),\Psi}x_{\mathcal{L}(\Psi)-1} + \mathcal{B}_{\mathcal{L}(\Psi),\Psi}) = x_L. \tag{2.17}$$

Moreover, note that (2.2) and (2.10) imply that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x = (x_0, x_1, \ldots, x_L) \in X_a$, $k \in \mathbb{N} \cap (0, \mathcal{L}(\Psi))$ it holds that

$$x_k = \mathfrak{M}_{a,\mathbb{D}_k(\Psi)}(\mathcal{W}_{k,\Psi}x_{k-1} + \mathcal{B}_{k,\Psi}) \tag{2.18}$$

This demonstrates that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x = (x_0, x_1, \ldots, x_L) \in X_a$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_a(\Psi))(x_0) = \mathcal{W}_{\mathcal{L}(\Psi),\Psi}x_{\mathcal{L}(\Psi)-1} + \mathcal{B}_{\mathcal{L}(\Psi),\Psi}. \tag{2.19}$$

Combining this with (2.17) establishes that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x = (x_0, x_1, \ldots, x_L) \in X_a$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_a(\Phi))\big((\mathcal{R}^{\mathbf{N}}_a(\Psi))(x_0)\big) = x_L = \big(\mathcal{R}^{\mathbf{N}}_a(\Phi \bullet \Psi)\big)(x_0). \tag{2.20}$$

This and (2.13) prove item (v). The proof of Proposition 2.1.2 is thus complete. □

2.1.3 Associativity of compositions of fully-connected feedforward ANNs

Lemma 2.1.3. *Let $\Phi_1, \Phi_2, \Phi_3 \in \mathbf{N}$ satisfy $\mathcal{I}(\Phi_1) = \mathcal{O}(\Phi_2)$, $\mathcal{I}(\Phi_2) = \mathcal{O}(\Phi_3)$, and $\mathcal{L}(\Phi_2) = 1$ (cf. Definition 1.3.1). Then*

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \tag{2.21}$$

(cf. Definition 2.1.1).

Proof of Lemma 2.1.3. Observe that the fact that for all $\Psi_1, \Psi_2 \in \mathbf{N}$ with $\mathcal{I}(\Psi_1) = \mathcal{O}(\Psi_2)$ it holds that $\mathcal{L}(\Psi_1 \bullet \Psi_2) = \mathcal{L}(\Psi_1) + \mathcal{L}(\Psi_2) - 1$ and the assumption that $\mathcal{L}(\Phi_2) = 1$ ensure that

$$\mathcal{L}(\Phi_1 \bullet \Phi_2) = \mathcal{L}(\Phi_1) \qquad \text{and} \qquad \mathcal{L}(\Phi_2 \bullet \Phi_3) = \mathcal{L}(\Phi_3) \tag{2.22}$$

(cf. Definition 2.1.1). Therefore, we obtain that

$$\mathcal{L}((\Phi_1 \bullet \Phi_2) \bullet \Phi_3) = \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_3) = \mathcal{L}(\Phi_1 \bullet (\Phi_2 \bullet \Phi_3)). \tag{2.23}$$

Next note that (2.22), (2.2), and the assumption that $\mathcal{L}(\Phi_2) = 1$ imply that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1)\}$ it holds that

$$(\mathcal{W}_{k,\Phi_1\bullet\Phi_2}, \mathcal{B}_{k,\Phi_1\bullet\Phi_2}) = \begin{cases} (\mathcal{W}_{1,\Phi_1}\mathcal{W}_{1,\Phi_2}, \mathcal{W}_{1,\Phi_1}\mathcal{B}_{1,\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k = 1 \\ (\mathcal{W}_{k,\Phi_1}, \mathcal{B}_{k,\Phi_1}) & : k > 1. \end{cases} \tag{2.24}$$

This, (2.2), and (2.23) prove that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_3) - 1\}$ it holds that

$$\begin{aligned} &(\mathcal{W}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}, \mathcal{B}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}) \\ &= \begin{cases} (\mathcal{W}_{k,\Phi_3}, \mathcal{B}_{k,\Phi_3}) & : k < \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{1,\Phi_1\bullet\Phi_2}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3}, \mathcal{W}_{1,\Phi_1\bullet\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_1\bullet\Phi_2}) & : k = \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1\bullet\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1\bullet\Phi_2}) & : k > \mathcal{L}(\Phi_3) \end{cases} \\ &= \begin{cases} (\mathcal{W}_{k,\Phi_3}, \mathcal{B}_{k,\Phi_3}) & : k < \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{1,\Phi_1\bullet\Phi_2}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3}, \mathcal{W}_{1,\Phi_1\bullet\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_1\bullet\Phi_2}) & : k = \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1}) & : k > \mathcal{L}(\Phi_3). \end{cases} \end{aligned} \tag{2.25}$$

Furthermore, observe that (2.2), (2.22), and (2.23) show that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_3) - 1\}$ it holds that

$$\begin{aligned} &(\mathcal{W}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}, \mathcal{B}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}) \\ &= \begin{cases} (\mathcal{W}_{k,\Phi_2\bullet\Phi_3}, \mathcal{B}_{k,\Phi_2\bullet\Phi_3}) & : k < \mathcal{L}(\Phi_2 \bullet \Phi_3) \\ (\mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_2\bullet\Phi_3),\Phi_2\bullet\Phi_3}, \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Phi_2\bullet\Phi_3),\Phi_2\bullet\Phi_3} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2 \bullet \Phi_3) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_2\bullet\Phi_3)+1,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_2\bullet\Phi_3)+1,\Phi_1}) & : k > \mathcal{L}(\Phi_2 \bullet \Phi_3) \end{cases} \\ &= \begin{cases} (\mathcal{W}_{k,\Phi_3}, \mathcal{B}_{k,\Phi_3}) & : k < \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_2\bullet\Phi_3}, \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_2\bullet\Phi_3} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1}) & : k > \mathcal{L}(\Phi_3). \end{cases} \end{aligned} \tag{2.26}$$

Combining this with (2.25) establishes that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1)+\mathcal{L}(\Phi_3)-1\}\backslash\{\mathcal{L}(\Phi_3)\}$ it holds that

$$(\mathcal{W}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}, \mathcal{B}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}) = (\mathcal{W}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}, \mathcal{B}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}). \tag{2.27}$$

Moreover, note that (2.24) and (2.2) ensure that

$$\mathcal{W}_{1,\Phi_1\bullet\Phi_2}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3} = \mathcal{W}_{1,\Phi_1}\mathcal{W}_{1,\Phi_2}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3} = \mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_2\bullet\Phi_3}. \tag{2.28}$$

In addition, observe that (2.24) and (2.2) demonstrate that

$$\begin{aligned} \mathcal{W}_{1,\Phi_1\bullet\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_1\bullet\Phi_2} &= \mathcal{W}_{1,\Phi_1}\mathcal{W}_{1,\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{W}_{1,\Phi_1}\mathcal{B}_{1,\Phi_2} + \mathcal{B}_{1,\Phi_1} \\ &= \mathcal{W}_{1,\Phi_1}(\mathcal{W}_{1,\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_2}) + \mathcal{B}_{1,\Phi_1} \\ &= \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_2\bullet\Phi_3} + \mathcal{B}_{1,\Phi_1}. \end{aligned} \tag{2.29}$$

Combining this and (2.28) with (2.27) proves that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_3) - 1\}$ it holds that

$$(\mathcal{W}_{k,(\Phi_1 \bullet \Phi_2) \bullet \Phi_3}, \mathcal{B}_{k,(\Phi_1 \bullet \Phi_2) \bullet \Phi_3}) = (\mathcal{W}_{k,\Phi_1 \bullet (\Phi_2 \bullet \Phi_3)}, \mathcal{B}_{k,\Phi_1 \bullet (\Phi_2 \bullet \Phi_3)}). \tag{2.30}$$

This and (2.23) imply that

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3). \tag{2.31}$$

The proof of Lemma 2.1.3 is thus complete. □

Lemma 2.1.4. *Let $\Phi_1, \Phi_2, \Phi_3 \in \mathbf{N}$ satisfy $\mathcal{I}(\Phi_1) = \mathcal{O}(\Phi_2)$, $\mathcal{I}(\Phi_2) = \mathcal{O}(\Phi_3)$, and $\mathcal{L}(\Phi_2) > 1$ (cf. Definition 1.3.1). Then*

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \tag{2.32}$$

(cf. Definition 2.1.1).

Proof of Lemma 2.1.4. Note that the fact that for all $\Psi, \Theta \in \mathbf{N}$ it holds that $\mathcal{L}(\Psi \bullet \Theta) = \mathcal{L}(\Psi) + \mathcal{L}(\Theta) - 1$ ensures that

$$\begin{aligned} \mathcal{L}((\Phi_1 \bullet \Phi_2) \bullet \Phi_3) &= \mathcal{L}(\Phi_1 \bullet \Phi_2) + \mathcal{L}(\Phi_3) - 1 \\ &= \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 2 \\ &= \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_2 \bullet \Phi_3) - 1 \\ &= \mathcal{L}(\Phi_1 \bullet (\Phi_2 \bullet \Phi_3)) \end{aligned} \tag{2.33}$$

(cf. Definition 2.1.1). Furthermore, observe that (2.2) shows that for all $k \in \{1, 2, \ldots, \mathcal{L}((\Phi_1 \bullet \Phi_2) \bullet \Phi_3)\}$ it holds that

$$\begin{aligned} &(\mathcal{W}_{k,(\Phi_1 \bullet \Phi_2) \bullet \Phi_3}, \mathcal{B}_{k,(\Phi_1 \bullet \Phi_2) \bullet \Phi_3}) \\ &= \begin{cases} (\mathcal{W}_{k,\Phi_3}, \mathcal{B}_{k,\Phi_3}) & : k < \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{1,\Phi_1 \bullet \Phi_2} \mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3}, \mathcal{W}_{1,\Phi_1 \bullet \Phi_2} \mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_1 \bullet \Phi_2}) & : k = \mathcal{L}(\Phi_3) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1 \bullet \Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1 \bullet \Phi_2}) & : k > \mathcal{L}(\Phi_3). \end{cases} \end{aligned} \tag{2.34}$$

Moreover, note that (2.2) and the assumption that $\mathcal{L}(\Phi_2) > 1$ ensure that for all $k \in \mathbb{N} \cap (\mathcal{L}(\Phi_3), \mathcal{L}((\Phi_1 \bullet \Phi_2) \bullet \Phi_3)]$ it holds that

$$\begin{aligned} &(\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1 \bullet \Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1 \bullet \Phi_2}) \\ &= \begin{cases} (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}) & : k - \mathcal{L}(\Phi_3) + 1 < \mathcal{L}(\Phi_2) \\ (\mathcal{W}_{1,\Phi_1} \mathcal{W}_{\mathcal{L}(\Phi_2),\Phi_2}, \mathcal{W}_{1,\Phi_1} \mathcal{B}_{\mathcal{L}(\Phi_2),\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k - \mathcal{L}(\Phi_3) + 1 = \mathcal{L}(\Phi_2) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1-\mathcal{L}(\Phi_2)+1,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1-\mathcal{L}(\Phi_2)+1,\Phi_1}) & : k - \mathcal{L}(\Phi_3) + 1 > \mathcal{L}(\Phi_2) \end{cases} \\ &= \begin{cases} (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}) & : k < \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\ (\mathcal{W}_{1,\Phi_1} \mathcal{W}_{\mathcal{L}(\Phi_2),\Phi_2}, \mathcal{W}_{1,\Phi_1} \mathcal{B}_{\mathcal{L}(\Phi_2),\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)-\mathcal{L}(\Phi_2)+2,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)-\mathcal{L}(\Phi_2)+2,\Phi_1}) & : k > \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1. \end{cases} \end{aligned} \tag{2.35}$$

Combining this with (2.34) proves that for all $k \in \{1, 2, \ldots, \mathcal{L}((\Phi_1 \bullet \Phi_2) \bullet \Phi_3)\}$ it holds that

$$\begin{aligned}
&(\mathcal{W}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}, \mathcal{B}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}) \\
&= \begin{cases}
(\mathcal{W}_{k,\Phi_3}, \mathcal{B}_{k,\Phi_3}) & : k < \mathcal{L}(\Phi_3) \\
(\mathcal{W}_{1,\Phi_2}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3}, \mathcal{W}_{1,\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_2}) & : k = \mathcal{L}(\Phi_3) \\
(\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}) & : \mathcal{L}(\Phi_3) < k < \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\
(\mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_2),\Phi_2}, \mathcal{W}_{1,\Phi_1}\mathcal{B}_{\mathcal{L}(\Phi_2),\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\
(\mathcal{W}_{k-\mathcal{L}(\Phi_3)-\mathcal{L}(\Phi_2)+2,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)-\mathcal{L}(\Phi_2)+2,\Phi_1}) & : k > \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1.
\end{cases}
\end{aligned} \tag{2.36}$$

In addition, observe that (2.2), the fact that $\mathcal{L}(\Phi_2 \bullet \Phi_3) = \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1$, and the assumption that $\mathcal{L}(\Phi_2) > 1$ demonstrate that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1 \bullet (\Phi_2 \bullet \Phi_3))\}$ it holds that

$$\begin{aligned}
&(\mathcal{W}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}, \mathcal{B}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}) \\
&= \begin{cases}
(\mathcal{W}_{k,\Phi_2\bullet\Phi_3}, \mathcal{B}_{k,\Phi_2\bullet\Phi_3}) & : k < \mathcal{L}(\Phi_2 \bullet \Phi_3) \\
(\mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_2\bullet\Phi_3),\Phi_2\bullet\Phi_3}, \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Phi_2\bullet\Phi_3),\Phi_2\bullet\Phi_3} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2 \bullet \Phi_3) \\
(\mathcal{W}_{k-\mathcal{L}(\Phi_2\bullet\Phi_3)+1,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_2\bullet\Phi_3)+1,\Phi_1}) & : k > \mathcal{L}(\Phi_2 \bullet \Phi_3)
\end{cases} \\
&= \begin{cases}
(\mathcal{W}_{k,\Phi_2\bullet\Phi_3}, \mathcal{B}_{k,\Phi_2\bullet\Phi_3}) & : k < \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\
(\mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_2)+\mathcal{L}(\Phi_3)-1,\Phi_2\bullet\Phi_3}, & \\
\quad \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Phi_2)+\mathcal{L}(\Phi_3)-1,\Phi_2\bullet\Phi_3} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\
(\mathcal{W}_{k-\mathcal{L}(\Phi_2)-\mathcal{L}(\Phi_3)+2,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_2)-\mathcal{L}(\Phi_3)+2,\Phi_1}) & : k > \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1
\end{cases} \\
&= \begin{cases}
(\mathcal{W}_{k,\Phi_3}, \mathcal{B}_{k,\Phi_3}) & : k < \mathcal{L}(\Phi_3) \\
(\mathcal{W}_{1,\Phi_2}\mathcal{W}_{\mathcal{L}(\Phi_3),\Phi_3}, \mathcal{W}_{1,\Phi_2}\mathcal{B}_{\mathcal{L}(\Phi_3),\Phi_3} + \mathcal{B}_{1,\Phi_2}) & : k = \mathcal{L}(\Phi_3) \\
(\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}) & : \mathcal{L}(\Phi_3) < k < \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\
(\mathcal{W}_{1,\Phi_1}\mathcal{W}_{\mathcal{L}(\Phi_2),\Phi_2}, \mathcal{W}_{1,\Phi}\mathcal{B}_{\mathcal{L}(\Phi_2),\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\
(\mathcal{W}_{k-\mathcal{L}(\Phi_2)-\mathcal{L}(\Phi_3)+2,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_2)-\mathcal{L}(\Phi_3)+2,\Phi_1}) & : k > \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1.
\end{cases}
\end{aligned} \tag{2.37}$$

This, (2.36), and (2.33) establish that for all $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1) + \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 2\}$ it holds that

$$(\mathcal{W}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}, \mathcal{B}_{k,(\Phi_1\bullet\Phi_2)\bullet\Phi_3}) = (\mathcal{W}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}, \mathcal{B}_{k,\Phi_1\bullet(\Phi_2\bullet\Phi_3)}). \tag{2.38}$$

Hence, we obtain that

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3). \tag{2.39}$$

The proof of Lemma 2.1.4 is thus complete. □

Corollary 2.1.5. *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*

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \tag{2.40}$$

(cf. Definition 2.1.1).

Proof of Corollary 2.1.5. Note that Lemma 2.1.3 and Lemma 2.1.4 establish (2.40). The proof of Corollary 2.1.5 is thus complete. □

2.1.4 Powers of fully-connected feedforward ANNs

Definition 2.1.6 (Powers of fully-connected feedforward ANNs)**.** *We denote by $(\cdot)^{\bullet n} \colon \{\Phi \in \mathbf{N} \colon \mathcal{I}(\Phi) = \mathcal{O}(\Phi)\} \to \mathbf{N}$, $n \in \mathbb{N}_0$, the functions which satisfy for all $n \in \mathbb{N}_0$, $\Phi \in \mathbf{N}$ with $\mathcal{I}(\Phi) = \mathcal{O}(\Phi)$ that*

$$\Phi^{\bullet n} = \begin{cases} \big(\mathrm{I}_{\mathcal{O}(\Phi)}, (0, 0, \ldots, 0)\big) \in \mathbb{R}^{\mathcal{O}(\Phi) \times \mathcal{O}(\Phi)} \times \mathbb{R}^{\mathcal{O}(\Phi)} & : n = 0 \\ \Phi \bullet (\Phi^{\bullet (n-1)}) & : n \in \mathbb{N} \end{cases} \tag{2.41}$$

(cf. Definitions 1.3.1, 1.5.5, and 2.1.1).

Lemma 2.1.7 (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*

$$\mathcal{H}(\Phi^{\bullet n}) = n\mathcal{H}(\Phi) \tag{2.42}$$

(cf. Definition 2.1.6).

Proof of Lemma 2.1.7. Observe that Proposition 2.1.2, (2.41), and induction establish (2.42). The proof of Lemma 2.1.7 is thus complete. □

2.2 Parallelizations of fully-connected feedforward ANNs

2.2.1 Parallelizations of fully-connected feedforward ANNs with the same length

Definition 2.2.1 (Parallelization of fully-connected feedforward ANNs)**.** *Let $n \in \mathbb{N}$.*

Then we denote by

$$\mathbf{P}_n \colon \big\{\Phi = (\Phi_1, \ldots, \Phi_n) \in \mathbf{N}^n \colon \mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \ldots = \mathcal{L}(\Phi_n)\big\} \to \mathbf{N} \tag{2.43}$$

the function which satisfies for all $\Phi = (\Phi_1, \ldots, \Phi_n) \in \mathbf{N}^n$, $k \in \{1, 2, \ldots, \mathcal{L}(\Phi_1)\}$ *with* $\mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \cdots = \mathcal{L}(\Phi_n)$ *that*

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

$$\text{and} \qquad \mathcal{B}_{k,\mathbf{P}_n(\Phi)} = \begin{pmatrix} \mathcal{B}_{k,\Phi_1} \\ \mathcal{B}_{k,\Phi_2} \\ \vdots \\ \mathcal{B}_{k,\Phi_n} \end{pmatrix} \tag{2.44}$$

(cf. Definition 1.3.1).

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

(i) it holds that

$$\mathbf{P}_n(\Phi) \in \left(\bigtimes_{k=1}^{L} \left(\mathbb{R}^{(\sum_{j=1}^{n} \mathbb{D}_k(\Phi_j)) \times (\sum_{j=1}^{n} \mathbb{D}_{k-1}(\Phi_j))} \times \mathbb{R}^{(\sum_{j=1}^{n} \mathbb{D}_k(\Phi_j))} \right) \right), \tag{2.45}$$

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

$$\mathbb{D}_k(\mathbf{P}_n(\Phi)) = \mathbb{D}_k(\Phi_1) + \mathbb{D}_k(\Phi_2) + \ldots + \mathbb{D}_k(\Phi_n), \tag{2.46}$$

and

(iii) it holds that

$$\mathcal{D}\big(\mathbf{P}_n(\Phi)\big) = \mathcal{D}(\Phi_1) + \mathcal{D}(\Phi_2) + \ldots + \mathcal{D}(\Phi_n) \tag{2.47}$$

(cf. Definition 2.2.1).

Proof of Lemma 2.2.2. Note that item (iii) in Lemma 1.3.3 and (2.44) imply that for all

$k \in \{1, 2, \ldots, L\}$ it holds that

$$\mathcal{W}_{k,\mathbf{P}_n(\Phi)} \in \mathbb{R}^{(\sum_{j=1}^{n} \mathbb{D}_k(\Phi_j)) \times (\sum_{j=1}^{n} \mathbb{D}_{k-1}(\Phi_j))} \qquad \text{and} \qquad \mathcal{B}_{k,\mathbf{P}_n(\Phi)} \in \mathbb{R}^{(\sum_{j=1}^{n} \mathbb{D}_{k-1}(\Phi_j))} \tag{2.48}$$

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

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

(i) it holds that

$$\mathcal{R}^{\mathbf{N}}_a(\mathbf{P}_n(\Phi)) \in C\big(\mathbb{R}^{[\sum_{j=1}^{n} \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^{n} \mathcal{O}(\Phi_j)]}\big) \tag{2.49}$$

and

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

$$\begin{aligned} &\big(\mathcal{R}^{\mathbf{N}}_a(\mathbf{P}_n(\Phi))\big)(x_1, x_2, \ldots, x_n) \\ &= \big((\mathcal{R}^{\mathbf{N}}_a(\Phi_1))(x_1), (\mathcal{R}^{\mathbf{N}}_a(\Phi_2))(x_2), \ldots, (\mathcal{R}^{\mathbf{N}}_a(\Phi_n))(x_n)\big) \in \mathbb{R}^{[\sum_{j=1}^{n} \mathcal{O}(\Phi_j)]} \end{aligned} \tag{2.50}$$

(cf. Definitions 1.3.4 and 2.2.1).

Proof of Proposition 2.2.3. Throughout this proof, let $L = \mathcal{L}(\Phi_1)$, for every $j \in \{1, 2, \ldots, n\}$ let

$$\begin{aligned} X^j = \big\{ &x = (x_0, x_1, \ldots, x_L) \in \mathbb{R}^{\mathbb{D}_0(\Phi_j)} \times \mathbb{R}^{\mathbb{D}_1(\Phi_j)} \times \cdots \times \mathbb{R}^{\mathbb{D}_L(\Phi_j)} : \\ &\big(\forall\, k \in \{1, 2, \ldots, L\} \colon x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), \mathbb{D}_k(\Phi_j)}(\mathcal{W}_{k,\Phi_j} x_{k-1} + \mathcal{B}_{k,\Phi_j})\big)\big\}, \end{aligned} \tag{2.51}$$

and let

$$\begin{aligned} \mathfrak{X} = \big\{ &\mathfrak{x} = (\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_L) \in \mathbb{R}^{\mathbb{D}_0(\mathbf{P}_n(\Phi))} \times \mathbb{R}^{\mathbb{D}_1(\mathbf{P}_n(\Phi))} \times \cdots \times \mathbb{R}^{\mathbb{D}_L(\mathbf{P}_n(\Phi))} : \\ &\big(\forall\, k \in \{1, 2, \ldots, L\} \colon \mathfrak{x}_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k) + \mathrm{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), \mathbb{D}_k(\mathbf{P}_n(\Phi))}(\mathcal{W}_{k,\mathbf{P}_n(\Phi)} \mathfrak{x}_{k-1} + \mathcal{B}_{k,\mathbf{P}_n(\Phi)})\big)\big\}. \end{aligned} \tag{2.52}$$

Observe that item (ii) in Lemma 2.2.2 and item (ii) in Lemma 1.3.3 imply that

$$\mathcal{I}(\mathbf{P}_n(\Phi)) = \mathbb{D}_0(\mathbf{P}_n(\Phi)) = \sum_{j=1}^{n} \mathbb{D}_0(\Phi_n) = \sum_{j=1}^{n} \mathcal{I}(\Phi_n). \tag{2.53}$$

Furthermore, note that item (ii) in Lemma 2.2.2 and item (ii) in Lemma 1.3.3 ensure that

$$\mathcal{O}(\mathbf{P}_n(\Phi)) = \mathbb{D}_{\mathcal{L}(\mathbf{P}_n(\Phi))}(\mathbf{P}_n(\Phi)) = \sum_{j=1}^{n} \mathbb{D}_{\mathcal{L}(\Phi_n)}(\Phi_n) = \sum_{j=1}^{n} \mathcal{O}(\Phi_n). \tag{2.54}$$

Observe that (2.44) and item (ii) in Lemma 2.2.2 show that for all $\mathscr{a} \in C(\mathbb{R}, \mathbb{R})$, $k \in \{1, 2, \ldots, L\}$, $x^1 \in \mathbb{R}^{\mathbb{D}_k(\Phi_1)}$, $x^2 \in \mathbb{R}^{\mathbb{D}_k(\Phi_2)}$, $\ldots$, $x^n \in \mathbb{R}^{\mathbb{D}_k(\Phi_n)}$, $\mathfrak{x} \in \mathbb{R}^{[\sum_{j=1}^{n} \mathbb{D}_k(\Phi_j)]}$ with $\mathfrak{x} = (x^1, x^2, \ldots, x^n)$ it holds that

$$
\begin{aligned}
&\mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\mathbf{P}_n(\Phi))}(\mathcal{W}_{k,\mathbf{P}_n(\Phi)}\mathfrak{x} + \mathcal{B}_{k,\mathbf{P}_n(\Phi)}) \\
&= \mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\mathbf{P}_n(\Phi))}\left(\begin{pmatrix} \mathcal{W}_{k,\Phi_1} & 0 & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{k,\Phi_2} & 0 & \cdots & 0 \\ 0 & 0 & \mathcal{W}_{k,\Phi_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathcal{W}_{k,\Phi_n} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \mathcal{B}_{k,\Phi_1} \\ \mathcal{B}_{k,\Phi_2} \\ \mathcal{B}_{k,\Phi_3} \\ \vdots \\ \mathcal{B}_{k,\Phi_n} \end{pmatrix}\right) \\
&= \mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\mathbf{P}_n(\Phi))}\left(\begin{pmatrix} \mathcal{W}_{k,\Phi_1}x_1 + \mathcal{B}_{k,\Phi_1} \\ \mathcal{W}_{k,\Phi_2}x_2 + \mathcal{B}_{k,\Phi_2} \\ \mathcal{W}_{k,\Phi_3}x_3 + \mathcal{B}_{k,\Phi_3} \\ \vdots \\ \mathcal{W}_{k,\Phi_n}x_n + \mathcal{B}_{k,\Phi_n} \end{pmatrix}\right) = \begin{pmatrix} \mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\Phi_1)}(\mathcal{W}_{k,\Phi_1}x_1 + \mathcal{B}_{k,\Phi_1}) \\ \mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\Phi_2)}(\mathcal{W}_{k,\Phi_2}x_2 + \mathcal{B}_{k,\Phi_2}) \\ \mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\Phi_3)}(\mathcal{W}_{k,\Phi_3}x_3 + \mathcal{B}_{k,\Phi_3}) \\ \vdots \\ \mathfrak{M}_{\mathscr{a},\mathbb{D}_k(\Phi_n)}(\mathcal{W}_{k,\Phi_n}x_n + \mathcal{B}_{k,\Phi_n}) \end{pmatrix}.
\end{aligned} \tag{2.55}
$$

This proves that for all $k \in \{1, 2, \ldots, L\}$, $\mathfrak{x} = (\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_L) \in \mathfrak{X}$, $x^1 = (x_0^1, x_1^1, \ldots, x_L^1) \in X^1$, $x^2 = (x_0^2, x_1^2, \ldots, x_L^2) \in X^2$, $\ldots$, $x^n = (x_0^n, x_1^n, \ldots, x_L^n) \in X^n$ with $\mathfrak{x}_{k-1} = (x_{k-1}^1, x_{k-1}^2, \ldots, x_{k-1}^n)$ it holds that

$$
\mathfrak{x}_k = (x_k^1, x_k^2, \ldots, x_k^n). \tag{2.56}
$$

Induction, and (1.92) hence demonstrate that for all $k \in \{1, 2, \ldots, L\}$, $\mathfrak{x} = (\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_L) \in \mathfrak{X}$, $x^1 = (x_0^1, x_1^1, \ldots, x_L^1) \in X^1$, $x^2 = (x_0^2, x_1^2, \ldots, x_L^2) \in X^2$, $\ldots$, $x^n = (x_0^n, x_1^n, \ldots, x_L^n) \in X^n$ with $\mathfrak{x}_0 = (x_0^1, x_0^2, \ldots, x_0^n)$ it holds that

$$
\begin{aligned}
\big(\mathcal{R}_a^{\mathbf{N}}(\mathbf{P}_n(\Phi))\big)(\mathfrak{x}_0) &= \mathfrak{x}_L = (x_L^1, x_L^2, \ldots, x_L^n) \\
&= \big((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_0^1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_0^2), \ldots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_0^n)\big).
\end{aligned} \tag{2.57}
$$

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

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

$$
\mathcal{P}\big(\mathbf{P}_n(\Phi_1, \Phi_2, \ldots, \Phi_n)\big) \leq \tfrac{1}{2}\big[\textstyle\sum_{j=1}^{n} \mathcal{P}(\Phi_j)\big]^2 \tag{2.58}
$$

(cf. Definition 2.2.1).

Proof of Proposition 2.2.4. Throughout this proof, for every $j \in \{1, 2, \ldots, n\}$, $k \in \{0, 1, \ldots, L\}$ let $l_{j,k} = \mathbb{D}_k(\Phi_j)$. Note that item (ii) in Lemma 2.2.2 demonstrates that

$$\begin{aligned}
\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \ldots, \Phi_n)) &= \sum_{k=1}^{L} \Big[\textstyle\sum_{i=1}^{n} l_{i,k}\Big] \Big[\big(\textstyle\sum_{i=1}^{n} l_{i,k-1}\big) + 1\Big] \\
&= \sum_{k=1}^{L} \Big[\textstyle\sum_{i=1}^{n} l_{i,k}\Big] \Big[\big(\textstyle\sum_{j=1}^{n} l_{j,k-1}\big) + 1\Big] \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{L} l_{i,k}(l_{j,k-1} + 1) \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k,\ell=1}^{L} l_{i,k}(l_{j,\ell-1} + 1) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \Big[\textstyle\sum_{k=1}^{L} l_{i,k}\Big] \Big[\textstyle\sum_{\ell=1}^{L} (l_{j,\ell-1} + 1)\Big] \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{n} \Big[\textstyle\sum_{k=1}^{L} \frac{1}{2} l_{i,k}(l_{i,k-1} + 1)\Big] \Big[\textstyle\sum_{\ell=1}^{L} l_{j,\ell}(l_{j,\ell-1} + 1)\Big] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \tfrac{1}{2} \mathcal{P}(\Phi_i)\mathcal{P}(\Phi_j) = \tfrac{1}{2} \Big[\textstyle\sum_{i=1}^{n} \mathcal{P}(\Phi_i)\Big]^2.
\end{aligned} \tag{2.59}$$

The proof of Proposition 2.2.4 is thus complete. □

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

$$\big[\tfrac{n^2}{2}\big]\mathcal{P}(\Phi_1) \leq \big[\tfrac{n^2+n}{2}\big]\mathcal{P}(\Phi_1) \leq \mathcal{P}(\mathbf{P}_n(\Phi)) \leq n^2\mathcal{P}(\Phi_1) \leq \tfrac{1}{2}\big[\textstyle\sum_{i=1}^{n} \mathcal{P}(\Phi_i)\big]^2 \tag{2.60}$$

(cf. Definition 2.2.1).

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

$$\mathcal{D}(\Phi_1) = (l_0, l_1, \ldots, l_L). \tag{2.61}$$

Observe that (2.61) and the assumption that $\mathcal{D}(\Phi_1) = \mathcal{D}(\Phi_2) = \ldots = \mathcal{D}(\Phi_n)$ imply that for all $j \in \{1, 2, \ldots, n\}$ it holds that

$$\mathcal{D}(\Phi_j) = (l_0, l_1, \ldots, l_L). \tag{2.62}$$

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

$$\mathcal{P}(\mathbf{P}_n(\Phi)) = \sum_{j=1}^{L} (nl_j)\big((nl_{j-1}) + 1\big). \tag{2.63}$$

Hence, we obtain that

$$\mathcal{P}(\mathbf{P}_n(\Phi)) \leq \sum_{j=1}^{L}(nl_j)\big((nl_{j-1})+n\big) = n^2\left[\sum_{j=1}^{L} l_j(l_{j-1}+1)\right] = n^2\mathcal{P}(\Phi_1). \tag{2.64}$$

Furthermore, note that the assumption that $\mathcal{D}(\Phi_1) = \mathcal{D}(\Phi_2) = \ldots = \mathcal{D}(\Phi_n)$ and the fact that $\mathcal{P}(\Phi_1) \geq l_1(l_0+1) \geq 2$ ensure that

$$n^2\mathcal{P}(\Phi_1) \leq \tfrac{n^2}{2}[\mathcal{P}(\Phi_1)]^2 = \tfrac{1}{2}[n\mathcal{P}(\Phi_1)]^2 = \tfrac{1}{2}\left[\sum_{i=1}^{n}\mathcal{P}(\Phi_1)\right]^2 = \tfrac{1}{2}\left[\sum_{i=1}^{n}\mathcal{P}(\Phi_i)\right]^2. \tag{2.65}$$

Moreover, observe that (2.63) and the fact that for all $a, b \in \mathbb{N}$ it holds that

$$2(ab+1) = ab+1+(a-1)(b-1)+a+b \geq ab+a+b+1 = (a+1)(b+1) \tag{2.66}$$

show that

$$\begin{aligned}\mathcal{P}(\mathbf{P}_n(\Phi)) &\geq \tfrac{1}{2}\left[\sum_{j=1}^{L}(nl_j)(n+1)(l_{j-1}+1)\right]\\ &= \tfrac{n(n+1)}{2}\left[\sum_{j=1}^{L} l_j(l_{j-1}+1)\right] = \left[\tfrac{n^2+n}{2}\right]\mathcal{P}(\Phi_1).\end{aligned} \tag{2.67}$$

This, (2.64), and (2.65) establish (2.60). The proof of Corollary 2.2.5 is thus complete. □

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

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \ldots, \Phi_n)) \leq n\big[\textstyle\sum_{i=1}^{n}\mathcal{P}(\Phi_i)\big]. \tag{2.68}$$

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

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \ldots, \Phi_n)) \leq n^2\mathcal{P}(\Phi_1). \tag{2.69}$$

2.2.2 Representations of the identities with ReLU activation functions

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

$$\mathfrak{I}_1 = \left(\left(\begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}\right), \begin{pmatrix}1 & -1\end{pmatrix}, 0\right) \in \big((\mathbb{R}^{2\times 1}\times\mathbb{R}^2)\times(\mathbb{R}^{1\times 2}\times\mathbb{R}^1)\big) \tag{2.70}$$

and

$$\mathfrak{I}_d = \mathbf{P}_d(\mathfrak{I}_1, \mathfrak{I}_1, \ldots, \mathfrak{I}_1) \tag{2.71}$$

(cf. Definitions 1.3.1 and 2.2.1).

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

(i) it holds that

$$\mathcal{D}(\mathfrak{I}_d) = (d, 2d, d) \in \mathbb{N}^3 \tag{2.72}$$

and

(ii) it holds that

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_d) = \mathrm{id}_{\mathbb{R}^d} \tag{2.73}$$

(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 (2.70) shows that

$$\mathcal{D}(\mathfrak{I}_1) = (1, 2, 1) = (l_0, l_1, l_2). \tag{2.74}$$

This, (2.71), and Lemma 2.2.2 prove that

$$\mathcal{D}(\mathfrak{I}_d) = (d, 2d, d) \in \mathbb{N}^3. \tag{2.75}$$

This establishes item (i). Next note that (2.70) assures that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_1))(x) = \mathfrak{r}(x) - \mathfrak{r}(-x) = \max\{x, 0\} - \max\{-x, 0\} = x. \tag{2.76}$$

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

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_d))(x) &= \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}\big(\mathbf{P}_d(\mathfrak{I}_1, \mathfrak{I}_1, \ldots, \mathfrak{I}_1)\big)\big)(x_1, x_2, \ldots, x_d) \\ &= \big((\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_1))(x_1), (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_1))(x_2), \ldots, (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_1))(x_d)\big) \\ &= (x_1, x_2, \ldots, x_d) = x\end{aligned} \tag{2.77}$$

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

Lemma 2.2.8 (Fully-connected feedforward softplus identity ANNs)**.** *Let $d \in \mathbb{N}$ and let a be the softplus activation function (cf. Definition 1.2.11). Then*

$$\mathcal{R}^{\mathbf{N}}_{a}(\mathfrak{I}_d) = \mathrm{id}_{\mathbb{R}^d} \tag{2.78}$$

(cf. Definitions 1.3.4 and 2.2.6).

Proof of Lemma 2.2.8. Note that (1.47) and (2.70) ensure that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned}(\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_1))(x) &= \ln(1+\exp(x+0)) - \ln(1+\exp(-x+0)) + 0 \\ &= \ln(1+\exp(x)) - \ln(1+\exp(-x)) \\ &= \ln\left(\tfrac{1+\exp(x)}{1+\exp(-x)}\right) \\ &= \ln\left(\tfrac{\exp(x)(1+\exp(-x))}{1+\exp(-x)}\right) \\ &= \ln(\exp(x)) = x\end{aligned} \tag{2.79}$$

(cf. Definitions 1.3.4 and 2.2.6). Combining this and Proposition 2.2.3 demonstrates that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_d) \in C(\mathbb{R}^d, \mathbb{R}^d)$ and

$$\begin{aligned}(\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_d))(x) &= \big(\mathcal{R}_a^{\mathbf{N}}\big(\mathbf{P}_d(\mathfrak{I}_1, \mathfrak{I}_1, \ldots, \mathfrak{I}_1)\big)\big)(x_1, x_2, \ldots, x_d) \\ &= \big((\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_1))(x_2), \ldots, (\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_1))(x_d)\big) \\ &= (x_1, x_2, \ldots, x_d) = x\end{aligned} \tag{2.80}$$

(cf. Definition 2.2.1). The proof of Lemma 2.2.8 is thus complete. □

2.2.3 Extensions of fully-connected feedforward ANNs

Definition 2.2.9 (Extensions of fully-connected feedforward ANNs)**.** *Let $L \in \mathbb{N}$, $\mathbb{I} \in \mathbf{N}$ satisfy $\mathcal{I}(\mathbb{I}) = \mathcal{O}(\mathbb{I})$. Then we denote by*

$$\mathcal{E}_{L,\mathbb{I}} \colon \big\{\Phi \in \mathbf{N} \colon \big(\mathcal{L}(\Phi) \leq L \text{ and } \mathcal{O}(\Phi) = \mathcal{I}(\mathbb{I})\big)\big\} \to \mathbf{N} \tag{2.81}$$

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

$$\mathcal{E}_{L,\mathbb{I}}(\Phi) = (\mathbb{I}^{\bullet(L-\mathcal{L}(\Phi))}) \bullet \Phi \tag{2.82}$$

(cf. Definitions 1.3.1, 2.1.1, and 2.1.6).

Lemma 2.2.10 (Length of extensions of fully-connected feedforward ANNs)**.** *Let $d, \mathfrak{i} \in \mathbb{N}$, $\Psi \in \mathbf{N}$ satisfy $\mathcal{D}(\Psi) = (d, \mathfrak{i}, d)$ (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

$$\mathcal{D}(\Psi^{\bullet n}) = \begin{cases} (d, d) & : n = 0 \\ (d, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) & : n \in \mathbb{N} \end{cases} \tag{2.83}$$

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*

$$\mathcal{L}(\mathcal{E}_{L,\Psi}(\Phi)) = L \tag{2.84}$$

(cf. Definitions 2.1.6 and 2.2.9).

Proof of Lemma 2.2.10. Throughout this proof, let $\Phi \in \mathbf{N}$ satisfy $\mathcal{O}(\Phi) = d$. Observe that Lemma 2.1.7 and the fact that $\mathcal{H}(\Psi) = 1$ prove that for all $n \in \mathbb{N}_0$ it holds that

$$\mathcal{H}(\Psi^{\bullet n}) = n\mathcal{H}(\Psi) = n \tag{2.85}$$

(cf. Definition 2.1.6). Combining this with (1.79) and Lemma 1.3.3 implies that

$$\mathcal{H}(\Psi^{\bullet n}) = n, \qquad \mathcal{L}(\Psi^{\bullet n}) = n + 1, \qquad \text{and} \qquad \mathcal{D}(\Psi^{\bullet n}) \in \mathbb{N}^{n+2}. \tag{2.86}$$

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

$$\mathbb{N}^{n+2} \ni \mathcal{D}(\Psi^{\bullet n}) = \begin{cases} (d, d) & : n = 0 \\ (d, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) & : n \in \mathbb{N}. \end{cases} \tag{2.87}$$

We now prove (2.87) by induction on $n \in \mathbb{N}_0$. Note that the fact that

$$\Psi^{\bullet 0} = (\mathrm{I}_d, 0) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \tag{2.88}$$

establishes (2.87) in the base case $n = 0$ (cf. Definition 1.5.5). For the induction step assume that there exists $n \in \mathbb{N}_0$ which satisfies

$$\mathbb{N}^{n+2} \ni \mathcal{D}(\Psi^{\bullet n}) = \begin{cases} (d, d) & : n = 0 \\ (d, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) & : n \in \mathbb{N}. \end{cases} \tag{2.89}$$

Note that (2.89), (2.41), (2.86), item (i) in Proposition 2.1.2, and the fact that $\mathcal{D}(\Psi) = (d, \mathfrak{i}, d) \in \mathbb{N}^3$ imply that

$$\mathcal{D}(\Psi^{\bullet(n+1)}) = \mathcal{D}(\Psi \bullet (\Psi^{\bullet n})) = (d, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) \in \mathbb{N}^{n+3} \tag{2.90}$$

(cf. Definition 2.1.1). Induction therefore proves (2.87). This and (2.86) establish item (i). Observe that (2.82), item (iii) in Proposition 2.1.2, (2.85), 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

$$\begin{aligned} \mathcal{H}\big(\mathcal{E}_{L,\Psi}(\Phi)\big) &= \mathcal{H}\big((\Psi^{\bullet(L-\mathcal{L}(\Phi))}) \bullet \Phi\big) = \mathcal{H}\big(\Psi^{\bullet(L-\mathcal{L}(\Phi))}\big) + \mathcal{H}(\Phi) \\ &= (L - \mathcal{L}(\Phi)) + \mathcal{H}(\Phi) = L - 1. \end{aligned} \tag{2.91}$$

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

$$\mathcal{L}\big(\mathcal{E}_{L,\Psi}(\Phi)\big) = \mathcal{H}\big(\mathcal{E}_{L,\Psi}(\Phi)\big) + 1 = L. \tag{2.92}$$

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

Lemma 2.2.11 (Realizations of extensions of fully-connected feedforward ANNs)**.** *Let $a \in C(\mathbb{R}, \mathbb{R})$, $\mathbb{I} \in \mathbf{N}$ satisfy $\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}) = \mathrm{id}_{\mathbb{R}^{\mathcal{I}(\mathbb{I})}}$ (cf. Definitions 1.3.1 and 1.3.4). Then*

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

$$\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}) = \mathrm{id}_{\mathbb{R}^{\mathcal{I}(\mathbb{I})}} \tag{2.93}$$

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

$$\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}}(\Phi)) = \mathcal{R}_a^{\mathbf{N}}(\Phi) \tag{2.94}$$

(cf. Definitions 2.1.6 and 2.2.9).

Proof of Lemma 2.2.11. 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

$$\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}) \in C(\mathbb{R}^d, \mathbb{R}^d) \qquad \text{and} \qquad \forall\, x \in \mathbb{R}^d \colon (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}))(x) = x. \tag{2.95}$$

We now prove (2.95) by induction on $n \in \mathbb{N}_0$. Note that (2.41) 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 $\forall\, x \in \mathbb{R}^d \colon (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet 0}))(x) = x$. This establishes (2.95) 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 $\forall\, x \in \mathbb{R}^d \colon (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}))(x) = x$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(n+1)}) = \mathcal{R}_a^{\mathbf{N}}(\mathbb{I} \bullet (\mathbb{I}^{\bullet n})) = (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I})) \circ (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n})) \in C(\mathbb{R}^d, \mathbb{R}^d) \tag{2.96}$$

and

$$\begin{aligned} \forall\, x \in \mathbb{R}^d \colon \big(\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(n+1)})\big)(x) &= \big([\mathcal{R}_a^{\mathbf{N}}(\mathbb{I})] \circ [\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n})]\big)(x) \\ &= (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}))\big(\big(\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n})\big)(x)\big) = (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}))(x) = x. \end{aligned} \tag{2.97}$$

Induction therefore proves (2.95). This establishes item (i). Note (2.82), item (v) in Proposition 2.1.2, item (i), and the fact that $\mathcal{I}(\mathbb{I}) = \mathcal{O}(\Phi)$ ensure that

$$\begin{aligned} &\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}}(\Phi)) = \mathcal{R}_a^{\mathbf{N}}((\mathbb{I}^{\bullet(L-\mathcal{L}(\Phi))}) \bullet \Phi) \\ &\in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\mathbb{I})}) = C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{I}(\mathbb{I})}) = C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)}) \end{aligned} \tag{2.98}$$

and

$$\begin{aligned} \forall\, x \in \mathbb{R}^{\mathcal{I}(\Phi)} \colon \big(\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}}(\Phi))\big)(x) &= \big(\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet(L-\mathcal{L}(\Phi))})\big)\big((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)\big) \\ &= (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x). \end{aligned} \tag{2.99}$$

This establishes item (ii). The proof of Lemma 2.2.11 is thus complete. □

Lemma 2.2.12 (Architectures of extensions of fully-connected feedforward ANNs). *Let $d, \mathfrak{i}, L, \mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \ldots, l_{L-1} \in \mathbb{N}$, $\Phi, \Psi \in \mathbf{N}$ satisfy*

$$\mathfrak{L} \geq L, \qquad \mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_{L-1}, d), \qquad \text{and} \qquad \mathcal{D}(\Psi) = (d, \mathfrak{i}, d) \tag{2.100}$$

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

$$\mathcal{D}(\mathcal{E}_{\mathfrak{L},\Psi}(\Phi)) = \begin{cases} (l_0, l_1, \ldots, l_{L-1}, d) & : \mathfrak{L} = L \\ (l_0, l_1, \ldots, l_{L-1}, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) & : \mathfrak{L} > L \end{cases} \tag{2.101}$$

(cf. Definition 2.2.9).

Proof of Lemma 2.2.12. Observe that item (i) in Lemma 2.2.10 shows that

$$\mathcal{H}(\Psi^{\bullet(\mathfrak{L}-L)})) = \mathfrak{L} - L, \qquad \mathcal{D}(\Psi^{\bullet(\mathfrak{L}-L)}) \in \mathbb{N}^{\mathfrak{L}-L+2}, \tag{2.102}$$

$$\text{and} \qquad \mathcal{D}(\Psi^{\bullet(\mathfrak{L}-L)}) = \begin{cases} (d, d) & : \mathfrak{L} = L \\ (d, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) & : \mathfrak{L} > L \end{cases} \tag{2.103}$$

(cf. Definition 2.1.6). Combining this with Proposition 2.1.2 ensures that

$$\mathcal{H}\big((\Psi^{\bullet(\mathfrak{L}-L)}) \bullet \Phi\big) = \mathcal{H}(\Psi^{\bullet(\mathfrak{L}-L)}) + \mathcal{H}(\Phi) = (\mathfrak{L} - L) + L - 1 = \mathfrak{L} - 1, \tag{2.104}$$

$$\mathcal{D}((\Psi^{\bullet(\mathfrak{L}-L)}) \bullet \Phi) \in \mathbb{N}^{\mathfrak{L}+1}, \tag{2.105}$$

$$\text{and} \qquad \mathcal{D}((\Psi^{\bullet(\mathfrak{L}-L)}) \bullet \Phi) = \begin{cases} (l_0, l_1, \ldots, l_{L-1}, d) & : \mathfrak{L} = L \\ (l_0, l_1, \ldots, l_{L-1}, \mathfrak{i}, \mathfrak{i}, \ldots, \mathfrak{i}, d) & : \mathfrak{L} > L. \end{cases} \tag{2.106}$$

This and (2.82) establish (2.101). The proof of Lemma 2.2.12 is thus complete. □

2.2.4 Parallelizations of fully-connected feedforward ANNs with different lengths

Definition 2.2.13 (Parallelization of fully-connected feedforward ANNs with different length). *Let $n \in \mathbb{N}$, $\Psi = (\Psi_1, \ldots, \Psi_n) \in \mathbf{N}^n$ satisfy for all $j \in \{1, 2, \ldots, n\}$ that*

$$\mathcal{H}(\Psi_j) = 1 \qquad \text{and} \qquad \mathcal{I}(\Psi_j) = \mathcal{O}(\Psi_j) \tag{2.107}$$

(cf. Definition 1.3.1). Then we denote by

$$\mathrm{P}_{n,\Psi} \colon \big\{\Phi = (\Phi_1, \ldots, \Phi_n) \in \mathbf{N}^n \colon \big(\forall\, j \in \{1, 2, \ldots, n\} \colon \mathcal{O}(\Phi_j) = \mathcal{I}(\Psi_j)\big)\big\} \to \mathbf{N} \tag{2.108}$$

the function which satisfies for all $\Phi = (\Phi_1, \ldots, \Phi_n) \in \mathbf{N}^n$ with $\forall\, j \in \{1, 2, \ldots, n\}$:

$\mathcal{O}(\Phi_j) = \mathcal{I}(\Psi_j)$ *that*

$$\mathrm{P}_{n,\Psi}(\Phi) = \mathbf{P}_n\big(\mathcal{E}_{\max_{k\in\{1,2,\ldots,n\}}\mathcal{L}(\Phi_k),\Psi_1}(\Phi_1), \ldots, \mathcal{E}_{\max_{k\in\{1,2,\ldots,n\}}\mathcal{L}(\Phi_k),\Psi_n}(\Phi_n)\big) \tag{2.109}$$

(cf. Definitions 2.2.1 and 2.2.9 and Lemma 2.2.10).

Lemma 2.2.14 (Realizations for parallelizations of fully-connected feedforward ANNs with different length)**.** *Let $a \in C(\mathbb{R},\mathbb{R})$, $n \in \mathbb{N}$, $\mathbb{I} = (\mathbb{I}_1, \ldots, \mathbb{I}_n)$, $\Phi = (\Phi_1, \ldots, \Phi_n) \in \mathbf{N}^n$ satisfy for all $j \in \{1,2,\ldots,n\}$, $x \in \mathbb{R}^{\mathcal{O}(\Phi_j)}$ that $\mathcal{H}(\mathbb{I}_j) = 1$, $\mathcal{I}(\mathbb{I}_j) = \mathcal{O}(\mathbb{I}_j) = \mathcal{O}(\Phi_j)$, and $(\mathcal{R}^{\mathbf{N}}_a(\mathbb{I}_j))(x) = x$ (cf. Definitions 1.3.1 and 1.3.4). Then*

(i) it holds that

$$\mathcal{R}^{\mathbf{N}}_a\big(\mathrm{P}_{n,\mathbb{I}}(\Phi)\big) \in C\big(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}\big) \tag{2.110}$$

and

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

$$\begin{aligned}&\big(\mathcal{R}^{\mathbf{N}}_a(\mathrm{P}_{n,\mathbb{I}}(\Phi))\big)(x_1, x_2, \ldots, x_n)\\ &= \big((\mathcal{R}^{\mathbf{N}}_a(\Phi_1))(x_1), (\mathcal{R}^{\mathbf{N}}_a(\Phi_2))(x_2), \ldots, (\mathcal{R}^{\mathbf{N}}_a(\Phi_n))(x_n)\big) \in \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}\end{aligned} \tag{2.111}$$

(cf. Definition 2.2.13).

Proof of Lemma 2.2.14. Throughout this proof, let $L \in \mathbb{N}$ satisfy $L = \max_{j\in\{1,2,\ldots,n\}} \mathcal{L}(\Phi_j)$. Note that item (ii) in Lemma 2.2.10, the assumption that for all $j \in \{1,2,\ldots,n\}$ it holds that $\mathcal{H}(\mathbb{I}_j) = 1$, (2.82), (2.4), and item (ii) in Lemma 2.2.11 demonstrate

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

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

$$\big(\mathcal{R}^{\mathbf{N}}_a(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j))\big)(x) = (\mathcal{R}^{\mathbf{N}}_a(\Phi_j))(x) \tag{2.112}$$

(cf. Definition 2.2.9). Items (i) and (ii) in Proposition 2.2.3 therefore imply

(A) that

$$\mathcal{R}^{\mathbf{N}}_a\big(\mathbf{P}_n\big(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \ldots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n)\big)\big) \in C\big(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}\big) \tag{2.113}$$

and

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

$$\begin{aligned}&\big(\mathcal{R}_a^{\mathbf{N}}\big(\mathbf{P}_n\big(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \ldots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n)\big)\big)\big)(x_1, x_2, \ldots, x_n)\\ &= \Big(\big(\mathcal{R}_a^{\mathbf{N}}\big(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1)\big)\big)(x_1), \big(\mathcal{R}_a^{\mathbf{N}}\big(\mathcal{E}_{L,\mathbb{I}_2}(\Phi_2)\big)\big)(x_2), \ldots, \big(\mathcal{R}_a^{\mathbf{N}}\big(\mathcal{E}_{L,\mathbb{I}_n}(\Phi_n)\big)\big)(x_n)\Big)\\ &= \Big((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \ldots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n)\Big)\end{aligned} \tag{2.114}$$

(cf. Definition 2.2.1). Combining this with (2.109) and the fact that $L = \max_{j\in\{1,2,\ldots,n\}} \mathcal{L}(\Phi_j)$ ensures

(C) that

$$\mathcal{R}_a^{\mathbf{N}}\big(\mathrm{P}_{n,\mathbb{I}}(\Phi)\big) \in C\big(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}\big) \tag{2.115}$$

and

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

$$\begin{aligned}&\big(\mathcal{R}_a^{\mathbf{N}}\big(\mathrm{P}_{n,\mathbb{I}}(\Phi)\big)\big)(x_1, x_2, \ldots, x_n)\\ &= \big(\mathcal{R}_a^{\mathbf{N}}\big(\mathbf{P}_n\big(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \ldots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n)\big)\big)\big)(x_1, x_2, \ldots, x_n)\\ &= \Big((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \ldots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n)\Big).\end{aligned} \tag{2.116}$$

This establishes items items (i) and (ii). The proof of Lemma 2.2.14 is thus complete. □

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

$$F_d(x) = (\max\{|x_1|\}, \max\{|x_1|, |x_2|\}, \ldots, \max\{|x_1|, |x_2|, \ldots, |x_d|\}). \tag{2.117}$$

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

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi) = F_d \tag{2.118}$$

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

2.3 Scalar multiplications of fully-connected feedforward ANNs

2.3.1 Affine transformations as fully-connected feedforward ANNs

Definition 2.3.1 (Fully-connected feedforward 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*

$$\mathbf{A}_{W,B} \in (\mathbb{R}^{m\times n} \times \mathbb{R}^m) \subseteq \mathbf{N} \tag{2.119}$$

the fully-connected feedforward ANN given by

$$\mathbf{A}_{W,B} = (W, B) \tag{2.120}$$

(cf. Definitions 1.3.1 and 1.3.2).

Lemma 2.3.2 (Realizations of fully-connected feedforward 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 $\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$, and

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

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B}))(x) = Wx + B \tag{2.121}$$

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

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

$$\mathcal{D}(\mathbf{A}_{W,B}) = (n, m) \in \mathbb{N}^2. \tag{2.122}$$

This establishes item (i). Furthermore, observe that the fact that

$$\mathbf{A}_{W,B} = (W, B) \in (\mathbb{R}^{m \times n} \times \mathbb{R}^m) \tag{2.123}$$

and (1.92) imply that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ it holds that $\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$ and

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B}))(x) = Wx + B. \tag{2.124}$$

This proves items (ii) and (iii). The proof of Lemma 2.3.2 is thus complete. □

Lemma 2.3.3 (Compositions with fully-connected feedforward 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

$$\mathcal{D}(\mathbf{A}_{W,B} \bullet \Phi) = (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \ldots, \mathbb{D}_{\mathcal{H}(\Phi)}(\Phi), m), \tag{2.125}$$

(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 $\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^m)$,

(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*

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B} \bullet \Phi))(x) = W\big((\mathcal{R}^{\mathbf{N}}_a(\Phi))(x)\big) + B, \tag{2.126}$$

(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*

$$\mathcal{D}(\Phi \bullet \mathbf{A}_{W,B}) = (n, \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \ldots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \tag{2.127}$$

(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* $\mathcal{R}^{\mathbf{N}}_a(\Phi \bullet \mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^{\mathcal{O}(\Phi)})$, *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*

$$(\mathcal{R}^{\mathbf{N}}_a(\Phi \bullet \mathbf{A}_{W,B}))(x) = (\mathcal{R}^{\mathbf{N}}_a(\Phi))(Wx + B) \tag{2.128}$$

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

Proof of Lemma 2.3.3. Note that Lemma 2.3.2 demonstrates 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 $\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$ and

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{W,B}))(x) = Wx + B \tag{2.129}$$

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

2.3.2 Scalar multiplications of fully-connected feedforward ANNs

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

$$\lambda \circledast \Phi = \mathbf{A}_{\lambda \mathrm{I}_{\mathcal{O}(\Phi)}, 0} \bullet \Phi \tag{2.130}$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, and 2.3.1).

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

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

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

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

$$\mathcal{R}^{\mathbf{N}}_a(\lambda \circledast \Phi) = \lambda\big[\mathcal{R}^{\mathbf{N}}_a(\Phi)\big] \tag{2.131}$$

(cf. Definitions 1.3.4 and 2.3.4).

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

$$L = \mathcal{L}(\Phi) \qquad \text{and} \qquad (l_0, l_1, \ldots, l_L) = \mathcal{D}(\Phi). \tag{2.132}$$

Observe that item (i) in Lemma 2.3.2 shows that

$$\mathcal{D}(\mathbf{A}_{\lambda \mathrm{I}_{\mathcal{O}(\Phi)},0}) = (\mathcal{O}(\Phi), \mathcal{O}(\Phi)) \tag{2.133}$$

(cf. Definitions 1.5.5 and 2.3.1). Combining this and item (i) in Lemma 2.3.3 ensures that

$$\mathcal{D}(\lambda \circledast \Phi) = \mathcal{D}(\mathbf{A}_{\lambda \mathrm{I}_{\mathcal{O}(\Phi)},0} \bullet \Phi) = (l_0, l_1, \ldots, l_{L-1}, \mathcal{O}(\Phi)) = \mathcal{D}(\Phi) \tag{2.134}$$

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

$$\begin{aligned} \big(\mathcal{R}^{\mathbf{N}}_a(\lambda \circledast \Phi)\big)(x) &= \big(\mathcal{R}^{\mathbf{N}}_a(\mathbf{A}_{\lambda \mathrm{I}_{\mathcal{O}(\Phi)},0} \bullet \Phi)\big)(x) \\ &= \lambda \, \mathrm{I}_{\mathcal{O}(\Phi)}\big((\mathcal{R}^{\mathbf{N}}_a(\Phi))(x)\big) \\ &= \lambda\big((\mathcal{R}^{\mathbf{N}}_a(\Phi))(x)\big) \end{aligned} \tag{2.135}$$

(cf. Definition 1.3.4). This establishes items (ii) and (iii). The proof of Lemma 2.3.5 is thus complete. □

2.4 Sums of fully-connected feedforward ANNs with the same length

2.4.1 Sums of vectors as fully-connected feedforward ANNs

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

$$\mathbb{S}_{m,n} \in (\mathbb{R}^{m \times (mn)} \times \mathbb{R}^m) \subseteq \mathbf{N} \tag{2.136}$$

the fully-connected feedforward ANN given by

$$\mathbb{S}_{m,n} = \mathbf{A}_{(\mathrm{I}_m \;\; \mathrm{I}_m \;\; \ldots \;\; \mathrm{I}_m),0} \tag{2.137}$$

(cf. Definitions 1.3.1, 1.3.2, 1.5.5, and 2.3.1).

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 $\mathcal{R}^{\mathbf{N}}_a(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$, and

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

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbb{S}_{m,n}))(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} x_k \tag{2.138}$$

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

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

$$\mathcal{D}(\mathbb{S}_{m,n}) = (mn, m) \in \mathbb{N}^2 \tag{2.139}$$

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

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_a(\mathbb{S}_{m,n}))(x_1, x_2, \ldots, x_n) &= \big(\mathcal{R}^{\mathbf{N}}_a\big(\mathbf{A}_{(\mathrm{I}_m \ \mathrm{I}_m \ \ldots \ \mathrm{I}_m),0}\big)\big)(x_1, x_2, \ldots, x_n) \\ &= (\mathrm{I}_m \ \ \mathrm{I}_m \ \ \ldots \ \ \mathrm{I}_m)(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} x_k\end{aligned} \tag{2.140}$$

(cf. Definitions 1.3.4, 1.5.5, and 2.3.1). This establishes items (ii) and (iii). The proof of Lemma 2.4.2 is thus complete. □

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

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

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

$$\big(\mathcal{R}^{\mathbf{N}}_a(\mathbb{S}_{m,n} \bullet \Phi)\big)(x) = \sum_{k=1}^{n} y_k \tag{2.141}$$

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

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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} x_k \tag{2.142}$$

(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 2.4.3 is thus complete. □

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

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

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

$$\big(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbb{S}_{\mathcal{I}(\Phi),n})\big)(x_1, x_2, \ldots, x_n) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))\left(\sum_{k=1}^{n} x_k\right) \tag{2.143}$$

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

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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} x_k \tag{2.144}$$

(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 2.4.4 is thus complete. □

2.4.2 Concatenation of vectors as fully-connected feedforward 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 fully-connected feedforward ANNs)**.** *Let $m, n \in \mathbb{N}$. Then we denote by*

$$\mathbb{T}_{m,n} \in (\mathbb{R}^{(mn)\times m} \times \mathbb{R}^{mn}) \subseteq \mathbf{N} \tag{2.145}$$

the fully-connected feedforward ANN given by

$$\mathbb{T}_{m,n} = \mathbf{A}_{(\mathrm{I}_m \ \mathrm{I}_m \ \ldots \ \mathrm{I}_m)^*, 0} \tag{2.146}$$

(cf. Definitions 1.3.1, 1.3.2, 1.5.5, 2.3.1, and 2.4.5).

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 $\mathcal{R}^{\mathbf{N}}_a(\mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$, and

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

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbb{T}_{m,n}))(x) = (x, x, \ldots, x) \tag{2.147}$$

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

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

$$\mathcal{D}(\mathbb{T}_{m,n}) = (m, mn) \in \mathbb{N}^2 \tag{2.148}$$

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

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_a(\mathbb{T}_{m,n}))(x) &= \big(\mathcal{R}^{\mathbf{N}}_a\big(\mathbf{A}_{(\mathrm{I}_m \;\; \mathrm{I}_m \;\; \ldots \;\; \mathrm{I}_m)^*, 0}\big)\big)(x) \\ &= (\mathrm{I}_m \;\; \mathrm{I}_m \;\; \ldots \;\; \mathrm{I}_m)^* x = (x, x, \ldots, x)\end{aligned} \tag{2.149}$$

(cf. Definitions 1.3.4, 1.5.5, 2.3.1, and 2.4.5). This establishes items (ii) and (iii). The proof of Lemma 2.4.7 is thus complete. □

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

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

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

$$\big(\mathcal{R}^{\mathbf{N}}_a(\mathbb{T}_{\mathcal{O}(\Phi),n} \bullet \Phi)\big)(x) = \big((\mathcal{R}^{\mathbf{N}}_a(\Phi))(x), (\mathcal{R}^{\mathbf{N}}_a(\Phi))(x), \ldots, (\mathcal{R}^{\mathbf{N}}_a(\Phi))(x)\big) \tag{2.150}$$

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

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

$$(\mathcal{R}^{\mathbf{N}}_a(\mathbb{T}_{m,n}))(x) = (x, x, \ldots, x) \tag{2.151}$$

(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 2.4.8 is thus complete. □

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

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

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

$$\big(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbb{T}_{m,n})\big)(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x, x, \ldots, x) \tag{2.152}$$

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

Proof of Lemma 2.4.9. Note that Lemma 2.4.7 implies 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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}))(x) = (x, x, \ldots, x) \tag{2.153}$$

(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 2.4.9 is thus complete. □

2.4.3 Sums of fully-connected feedforward ANNs

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

$$\mathcal{L}(\Phi_k) = \mathcal{L}(\Phi_m), \qquad \mathcal{I}(\Phi_k) = \mathcal{I}(\Phi_m), \qquad \text{and} \qquad \mathcal{O}(\Phi_k) = \mathcal{O}(\Phi_m) \tag{2.154}$$

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

$$\bigoplus_{k=m}^{n} \Phi_k = \big(\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet \big[\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)\big] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}\big) \in \mathbf{N} \tag{2.155}$$

(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 fully-connected feedforward ANNs)**.** *Let $m \in \mathbb{Z}$, $n \in \{m, m+1, \ldots\}$, $\Phi_m, \Phi_{m+1}, \ldots, \Phi_n \in \mathbf{N}$ satisfy for all $k \in \{m, m+1, \ldots, n\}$ that*

$$\mathcal{L}(\Phi_k) = \mathcal{L}(\Phi_m), \qquad \mathcal{I}(\Phi_k) = \mathcal{I}(\Phi_m), \qquad \text{and} \qquad \mathcal{O}(\Phi_k) = \mathcal{O}(\Phi_m) \tag{2.156}$$

(cf. Definition 1.3.1). Then

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

(ii) it holds that

$$\mathcal{D}\left(\bigoplus_{k=m}^{n} \Phi_k\right) = \left(\mathcal{I}(\Phi_m), \sum_{k=m}^{n} \mathbb{D}_1(\Phi_k), \sum_{k=m}^{n} \mathbb{D}_2(\Phi_k), \ldots, \sum_{k=m}^{n} \mathbb{D}_{\mathcal{H}(\Phi_m)}(\Phi_k), \mathcal{O}(\Phi_m)\right), \tag{2.157}$$

and

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

$$\mathcal{R}_a^{\mathbf{N}}\left(\bigoplus_{k=m}^{n} \Phi_k\right) = \sum_{k=m}^{n} (\mathcal{R}_a^{\mathbf{N}}(\Phi_k)) \tag{2.158}$$

(cf. Definitions 1.3.4 and 2.4.10).

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

$$\begin{aligned}
&\mathcal{D}\big(\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)\big) \\
&= \left(\sum_{k=m}^{n} \mathbb{D}_0(\Phi_k), \sum_{k=m}^{n} \mathbb{D}_1(\Phi_k), \ldots, \sum_{k=m}^{n} \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \sum_{k=m}^{n} \mathbb{D}_{\mathcal{L}(\Phi_m)}(\Phi_k)\right) \\
&= \left((n-m+1)\mathcal{I}(\Phi_m), \sum_{k=m}^{n} \mathbb{D}_1(\Phi_k), \sum_{k=m}^{n} \mathbb{D}_2(\Phi_k), \ldots, \sum_{k=m}^{n} \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k),\right. \\
&\qquad\qquad \left.(n-m+1)\mathcal{O}(\Phi_m)\right)
\end{aligned} \tag{2.159}$$

(cf. Definition 2.2.1). Furthermore, note that item (i) in Lemma 2.4.2 shows that

$$\mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m),n-m+1}) = ((n-m+1)\mathcal{O}(\Phi_m), \mathcal{O}(\Phi_m)) \tag{2.160}$$

(cf. Definition 2.4.1). This, (2.159), and item (i) in Proposition 2.1.2 ensure that

$$\begin{aligned}
&\mathcal{D}\big(\mathbb{S}_{\mathcal{O}(\Phi_m),n-m+1} \bullet \big[\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)\big]\big) \\
&= \left((n-m+1)\mathcal{I}(\Phi_m), \sum_{k=m}^{n} \mathbb{D}_1(\Phi_k), \sum_{k=m}^{n} \mathbb{D}_2(\Phi_k), \ldots, \sum_{k=m}^{n} \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m)\right).
\end{aligned} \tag{2.161}$$

Moreover, observe that item (i) in Lemma 2.4.7 proves that

$$\mathcal{D}\big(\mathbb{T}_{\mathcal{I}(\Phi_m),n-m+1}\big) = (\mathcal{I}(\Phi_m), (n-m+1)\mathcal{I}(\Phi_m)) \tag{2.162}$$

(cf. Definitions 2.1.1 and 2.4.6). Combining this, (2.161), and item (i) in Proposition 2.1.2

implies that

$$\begin{aligned}
&\mathcal{D}\Big(\bigoplus_{k=m}^{n} \Phi_k\Big) \\
&= \mathcal{D}\big(\mathbb{S}_{\mathcal{O}(\Phi_m),(n-m+1)} \bullet \big[\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)\big] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m),(n-m+1)}\big) \\
&= \Big(\mathcal{I}(\Phi_m), \sum_{k=m}^{n} \mathbb{D}_1(\Phi_k), \sum_{k=m}^{n} \mathbb{D}_2(\Phi_k), \ldots, \sum_{k=m}^{n} \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m)\Big)
\end{aligned} \tag{2.163}$$

(cf. Definition 2.4.10). This establishes items (i) and (ii). Note that Lemma 2.4.9 and (2.159) demonstrate that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi_m)}$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}\big([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m),n-m+1}\big) \in C(\mathbb{R}^{\mathcal{I}(\Phi_m)}, \mathbb{R}^{(n-m+1)\mathcal{O}(\Phi_m)}) \tag{2.164}$$

and

$$\begin{aligned}
&\big(\mathcal{R}_a^{\mathbf{N}}\big([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m),n-m+1}\big)\big)(x) \\
&= \big(\mathcal{R}_a^{\mathbf{N}}\big(\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)\big)\big)(x, x, \ldots, x)
\end{aligned} \tag{2.165}$$

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

$$\begin{aligned}
&\big(\mathcal{R}_a^{\mathbf{N}}\big([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m),n-m+1}\big)\big)(x) \\
&= \big((\mathcal{R}_a^{\mathbf{N}}(\Phi_m))(x), (\mathcal{R}_a^{\mathbf{N}}(\Phi_{m+1}))(x), \ldots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x)\big) \in \mathbb{R}^{(n-m+1)\mathcal{O}(\Phi_m)}.
\end{aligned} \tag{2.166}$$

Lemma 2.4.3, (2.160), and Corollary 2.1.5 hence ensure that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^{\mathcal{I}(\Phi_m)}$ it holds that $\mathcal{R}_a^{\mathbf{N}}\big(\bigoplus_{k=m}^{n} \Phi_k\big) \in C(\mathbb{R}^{\mathcal{I}(\Phi_m)}, \mathbb{R}^{\mathcal{O}(\Phi_m)})$ and

$$\begin{aligned}
&\Big(\mathcal{R}_a^{\mathbf{N}}\Big(\bigoplus_{k=m}^{n} \Phi_k\Big)\Big)(x) \\
&= \big(\mathcal{R}_a^{\mathbf{N}}\big(\mathbb{S}_{\mathcal{O}(\Phi_m),n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \ldots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m),n-m+1}\big)\big)(x) \\
&= \sum_{k=m}^{n} (\mathcal{R}_a^{\mathbf{N}}(\Phi_k))(x).
\end{aligned} \tag{2.167}$$

This proves item (iii). The proof of Lemma 2.4.11 is thus complete. □

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). ANN approximation results for multi-dimensional target functions are treated in Chapter 4 below.

In the scientific literature the capacity of ANNs to approximate certain classes of target functions has been thoroughly studied; cf., for instance, [14, 42, 91, 211, 212] for early universal ANN approximation results, cf., for example, [28, 44, 181, 347, 388, 437] and the references therein for more recent ANN approximation results establishing rates in the approximation of different classes of target functions, and cf., for instance, [133, 185, 269, 384] and the references therein for approximation capacities of ANNs related to solutions of PDEs (cf. also Chapters 16 and 17 in Part VI of these lecture notes for machine learning methods for PDEs). This chapter is based on Ackermann et al. [3, Section 4.2] (cf., for example, also Hutzenthaler et al. [217, Section 3.4]).

3.1 Linear interpolation of one-dimensional functions

3.1.1 On the modulus of continuity

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

$h \in [0, \infty]$ that

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

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 $f \colon A \to \mathbb{R}$ 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 $\limsup_{h \searrow 0} w_f(h) = 0$,

(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) implies items (i), (ii), (iii), and (iv). The proof of Lemma 3.1.2 is thus complete. □

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

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

(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 $[x - h, x + h] \cap [y - \mathfrak{h}, y + \mathfrak{h}] \cap [a, b] \neq \emptyset$ establishes 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

$$|x - z| \leq h \qquad \text{and} \qquad |y - z| \leq \mathfrak{h}. \tag{3.3}$$

Items (i) and (iv) in Lemma 3.1.2 therefore demonstrate 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

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

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

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

The proof of Lemma 3.1.3 is thus complete. □

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

$$|f(x) - f(y)| \leq L|x-y|, \tag{3.6}$$

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

$$w_f(h) \leq Lh \tag{3.7}$$

(cf. Definition 3.1.1).

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

$$\begin{aligned} w_f(h) &= \sup\big(\big[\cup_{x,y\in A, |x-y|\leq h}\{|f(x)-f(y)|\}\big] \cup \{0\}\big) \\ &\leq \sup\big(\big[\cup_{x,y\in A, |x-y|\leq h}\{L|x-y|\}\big] \cup \{0\}\big) \\ &\leq \sup(\{Lh, 0\}) = Lh \end{aligned} \tag{3.8}$$

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

3.1.2 Linear interpolation of one-dimensional functions

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

$$\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K} \colon \mathbb{R} \to \mathbb{R} \tag{3.9}$$

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

$$(\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(x) = f_0, \qquad (\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(z) = f_K, \tag{3.10}$$

and

$$(\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(y) = f_{k-1} + \big(\tfrac{y-\mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}\big)(f_k - f_{k-1}). \tag{3.11}$$

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

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

$$(\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(\mathfrak{x}_k) = f_k, \tag{3.12}$$

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

$$(\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(x) = f_{k-1} + \left(\tfrac{x-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)(f_k - f_{k-1}), \tag{3.13}$$

and

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

$$(\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(x) = \left(\tfrac{\mathfrak{x}_k-x}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)f_{k-1} + \left(\tfrac{x-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)f_k. \tag{3.14}$$

(cf. Definition 3.1.5).

Proof of Lemma 3.1.6. Note that (3.10) and (3.11) prove items (i) and (ii). Observe that item (ii) implies that for all $k \in \{1, 2, \ldots, K\}$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ it holds that

$$\begin{aligned}(\mathscr{L}^{f_0,f_1,\ldots,f_K}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(x) &= f_{k-1} + \left(\tfrac{x-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)(f_k - f_{k-1}) \\ &= \left[\left(\tfrac{\mathfrak{x}_k-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right) - \left(\tfrac{x-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)\right]f_{k-1} + \left(\tfrac{x-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)f_k \\ &= \left(\tfrac{\mathfrak{x}_k-x}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)f_{k-1} + \left(\tfrac{x-\mathfrak{x}_{k-1}}{\mathfrak{x}_k-\mathfrak{x}_{k-1}}\right)f_k.\end{aligned} \tag{3.15}$$

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

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

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

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

and

(ii) it holds that

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

(cf. Definitions 3.1.1 and 3.1.5).

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

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

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

$$\mathfrak{l}(x) = (\mathscr{L}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \ldots, f(\mathfrak{x}_K)}_{\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K})(x) \tag{3.19}$$

(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 $k \in \{1, 2, \ldots, K\}$, $x, y \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ with $x \neq y$ it holds that

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

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

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

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

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| \leq L|x - y|. \tag{3.22}$$

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

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| = 0 \leq L|x - y|, \tag{3.23}$$

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

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| = 0 \leq L|x - y|, \tag{3.24}$$

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

$$|\mathfrak{l}(x) - \mathfrak{l}(y)| \leq L|x - y|. \tag{3.25}$$

This establishes item (i). Note that item (iii) in Lemma 3.1.6 implies that for all $k \in \{1, 2, \ldots, K\}$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ it holds that

$$\begin{aligned} |\mathfrak{l}(x) - f(x)| &= \left| \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) f(\mathfrak{x}_{k-1}) + \left(\frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) f(\mathfrak{x}_k) - f(x) \right| \\ &= \left| \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (f(\mathfrak{x}_{k-1}) - f(x)) + \left(\frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (f(\mathfrak{x}_k) - f(x)) \right| \\ &\leq \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) |f(\mathfrak{x}_{k-1}) - f(x)| + \left(\frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) |f(\mathfrak{x}_k) - f(x)|. \end{aligned} \tag{3.26}$$

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

$$\begin{aligned} |\mathfrak{l}(x) - f(x)| &\leq w_f(|\mathfrak{x}_k - \mathfrak{x}_{k-1}|) \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} + \frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) \\ &= w_f(|\mathfrak{x}_k - \mathfrak{x}_{k-1}|) \leq w_f(\max\nolimits_{j \in \{1,2,\ldots,K\}} |\mathfrak{x}_j - \mathfrak{x}_{j-1}|). \end{aligned} \tag{3.27}$$

This proves item (ii). The proof of Proposition 3.1.7 is thus complete. □

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

$$|f(x) - f(y)| \leq L|x - y|. \tag{3.28}$$

Then

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

$$\left| (\mathscr{L}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \ldots, f(\mathfrak{x}_K)}_{\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K})(x) - (\mathscr{L}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \ldots, f(\mathfrak{x}_K)}_{\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K})(y) \right| \leq L|x - y| \tag{3.29}$$

and

(ii) it holds that

$$\sup_{x \in [\mathfrak{x}_0, \mathfrak{x}_K]} \left| (\mathscr{L}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \ldots, f(\mathfrak{x}_K)}_{\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K})(x) - f(x) \right| \leq L \left(\max_{k \in \{1,2,\ldots,K\}} |\mathfrak{x}_k - \mathfrak{x}_{k-1}| \right) \tag{3.30}$$

(cf. Definition 3.1.5).

Proof of Corollary 3.1.8. Observe that the assumption that for all $x, y \in [\mathfrak{x}_0, \mathfrak{x}_K]$ it holds that $|f(x) - f(y)| \leq L|x - y|$ shows that

$$0 \leq \frac{|f(\mathfrak{x}_K) - f(\mathfrak{x}_0)|}{(\mathfrak{x}_K - \mathfrak{x}_0)} \leq \frac{L|\mathfrak{x}_K - \mathfrak{x}_0|}{(\mathfrak{x}_K - \mathfrak{x}_0)} = L. \tag{3.31}$$

Combining this, Lemma 3.1.4, and the assumption that for all $x, y \in [\mathfrak{x}_0, \mathfrak{x}_K]$ it holds that $|f(x) - f(y)| \leq L|x - y|$ with item (i) in Proposition 3.1.7 ensures that for all $x, y \in \mathbb{R}$ it holds that

$$\begin{aligned} &\left|(\mathscr{L}^{f(\mathfrak{x}_0),f(\mathfrak{x}_1),\ldots,f(\mathfrak{x}_K)}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(x) - (\mathscr{L}^{f(\mathfrak{x}_0),f(\mathfrak{x}_1),\ldots,f(\mathfrak{x}_K)}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(y)\right| \\ &\leq \left(\max_{k\in\{1,2,\ldots,K\}}\left(\frac{L|\mathfrak{x}_k - \mathfrak{x}_{k-1}|}{|\mathfrak{x}_k - \mathfrak{x}_{k-1}|}\right)\right)|x - y| = L|x - y|. \end{aligned} \tag{3.32}$$

This establishes item (i). Note that the assumption that for all $x, y \in [\mathfrak{x}_0, \mathfrak{x}_K]$ 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

$$\begin{aligned} \sup_{x\in[\mathfrak{x}_0,\mathfrak{x}_K]}\left|(\mathscr{L}^{f(\mathfrak{x}_0),f(\mathfrak{x}_1),\ldots,f(\mathfrak{x}_K)}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K})(x) - f(x)\right| &\leq w_f\left(\max_{k\in\{1,2,\ldots,K\}}|\mathfrak{x}_k - \mathfrak{x}_{k-1}|\right) \\ &\leq L\left(\max_{k\in\{1,2,\ldots,K\}}|\mathfrak{x}_k - \mathfrak{x}_{k-1}|\right). \end{aligned} \tag{3.33}$$

This proves item (ii). The proof of Corollary 3.1.8 is thus complete. □

3.2 Linear interpolation with fully-connected feedforward ANNs

3.2.1 Activation functions as fully-connected feedforward ANNs

Definition 3.2.1 (Activation functions as fully-connected feedforward ANNs)**.** *Let $n \in \mathbb{N}$. Then we denote by*

$$\mathfrak{i}_n \in ((\mathbb{R}^{n\times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n\times n} \times \mathbb{R}^n)) \subseteq \mathbf{N} \tag{3.34}$$

the fully-connected feedforward ANN given by

$$\mathfrak{i}_n = ((\mathrm{I}_n, 0), (\mathrm{I}_n, 0)) \tag{3.35}$$

(cf. Definitions 1.3.1 and 1.5.5).

Lemma 3.2.2 (Realization functions of fully-connected feedforward activation ANNs)**.** *Let $n \in \mathbb{N}$. Then*

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

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

$$\mathcal{R}_a^{\mathbf{N}}(\mathfrak{i}_n) = \mathfrak{M}_{a,n} \tag{3.36}$$

(cf. Definitions 1.2.1, 1.3.1, 1.3.4, and 3.2.1).

Proof of Lemma 3.2.2. Observe that the fact that $\mathfrak{i}_n \in ((\mathbb{R}^{n\times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n\times n} \times \mathbb{R}^n)) \subseteq \mathbf{N}$ demonstrates that

$$\mathcal{D}(\mathfrak{i}_n) = (n, n, n) \in \mathbb{N}^3 \tag{3.37}$$

(cf. Definitions 1.3.1 and 3.2.1). This establishes item (i). Note that (1.92) and the fact that

$$\mathfrak{i}_n = ((\mathrm{I}_n, 0), (\mathrm{I}_n, 0)) \in ((\mathbb{R}^{n\times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n\times n} \times \mathbb{R}^n)) \tag{3.38}$$

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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathfrak{i}_n))(x) = \mathrm{I}_n(\mathfrak{M}_{a,n}(\mathrm{I}_n x + 0)) + 0 = \mathfrak{M}_{a,n}(x). \tag{3.39}$$

This proves item (ii). The proof of Lemma 3.2.2 is thus complete. □

Lemma 3.2.3 (Compositions of fully-connected feedforward activation ANNs with general fully-connected feedforward ANNs)**.** *Let $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then*

(i) it holds that

$$\begin{aligned}&\mathcal{D}(\mathfrak{i}_{\mathcal{O}(\Phi)} \bullet \Phi)\\ &= (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \ldots, \mathbb{D}_{\mathcal{L}(\Phi)-1}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)) \in \mathbb{N}^{\mathcal{L}(\Phi)+2},\end{aligned} \tag{3.40}$$

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

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

(iv) it holds that

$$\begin{aligned}&\mathcal{D}(\Phi \bullet \mathfrak{i}_{\mathcal{I}(\Phi)})\\ &= (\mathbb{D}_0(\Phi), \mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \ldots, \mathbb{D}_{\mathcal{L}(\Phi)-1}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)) \in \mathbb{N}^{\mathcal{L}(\Phi)+2},\end{aligned} \tag{3.41}$$

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

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

(cf. Definitions 1.2.1, 1.3.4, 2.1.1, and 3.2.1).

Proof of Lemma 3.2.3. Observe that Lemma 3.2.2 ensures that for all $n \in \mathbb{N}$, $a \in C(\mathbb{R}, \mathbb{R})$ it holds that

$$\mathcal{R}_a^{\mathbf{N}}(\mathfrak{i}_n) = \mathfrak{M}_{a,n} \tag{3.42}$$

(cf. Definitions 1.2.1, 1.3.4, and 3.2.1). Combining this and Proposition 2.1.2 establishes items (i), (ii), (iii), (iv), (v), and (vi). The proof of Lemma 3.2.3 is thus complete. □

3.2.2 Representations for ReLU ANNs with one hidden neuron

Lemma 3.2.4. *Let $\alpha, \beta, h \in \mathbb{R}$, $\mathbf{H} \in \mathbf{N}$ satisfy*

$$\mathbf{H} = h \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) \tag{3.43}$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, and 3.2.1). 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 $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{H}) \in C(\mathbb{R}, \mathbb{R})$, and

(iv) it holds for all $x \in \mathbb{R}$ that $(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{H}))(x) = h \max\{\alpha x + \beta, 0\}$

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Lemma 3.2.4. Note that Lemma 2.3.2 implies that

$$\mathbf{A}_{\alpha,\beta} = (\alpha, \beta), \qquad \mathcal{D}(\mathbf{A}_{\alpha,\beta}) = (1, 1) \in \mathbb{N}^2, \qquad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{A}_{\alpha,\beta}) \in C(\mathbb{R}, \mathbb{R}), \tag{3.44}$$

and $\forall\, x \in \mathbb{R}\colon (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{A}_{\alpha,\beta}))(x) = \alpha x + \beta$ (cf. Definitions 1.2.4 and 1.3.4). Proposition 2.1.2, Lemma 3.2.2, Lemma 3.2.3, (1.26), (1.92), and (2.2) therefore demonstrate that

$$\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta} = ((\alpha, \beta), (1, 0)), \ \mathcal{D}(\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) = (1, 1, 1) \in \mathbb{N}^3, \ \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) \in C(\mathbb{R}, \mathbb{R}), \tag{3.45}$$

$$\text{and} \qquad \forall\, x \in \mathbb{R}\colon (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta}))(x) = \mathfrak{r}(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{A}_{\alpha,\beta})(x)) = \max\{\alpha x + \beta, 0\}. \tag{3.46}$$

This, Lemma 2.3.5, and (2.130) show that

$$\mathbf{H} = h \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta}) = ((\alpha, \beta), (h, 0)), \quad \mathcal{D}(\mathbf{H}) = (1, 1, 1), \quad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{H}) \in C(\mathbb{R}, \mathbb{R}), \tag{3.47}$$

$$\text{and} \qquad (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{H}))(x) = h((\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{i}_1 \bullet \mathbf{A}_{\alpha,\beta}))(x)) = h \max\{\alpha x + \beta, 0\}. \tag{3.48}$$

This proves items (i), (ii), (iii), and (iv). The proof of Lemma 3.2.4 is thus complete. □

3.2.3 ReLU ANN representations for linear interpolations

Proposition 3.2.5 (ReLU ANN representations for linear interpolations). *Let $K \in \mathbb{N}$, $f_0, f_1, \ldots, f_K, \mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K \in \mathbb{R}$ satisfy $\mathfrak{x}_0 < \mathfrak{x}_1 < \ldots < \mathfrak{x}_K$ and let $\mathbf{F} \in \mathbf{N}$ satisfy*

$$\mathbf{F} = \mathbf{A}_{1,f_0} \bullet \left(\bigoplus_{k=0}^{K} \left(\left(\frac{(f_{\min\{k+1,K\}} - f_k)}{(\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_{\min\{k,K-1\}})} - \frac{(f_k - f_{\max\{k-1,0\}})}{(\mathfrak{x}_{\max\{k,1\}} - \mathfrak{x}_{\max\{k-1,0\}})} \right) \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k}) \right) \right) \tag{3.49}$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.1). Then

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

(ii) it holds that $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) = \mathscr{L}_{\mathfrak{x}_0,\mathfrak{x}_1,\ldots,\mathfrak{x}_K}^{f_0,f_1,\ldots,f_K}$, 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 3.2.5. Throughout this proof, let $c_0, c_1, \ldots, c_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \ldots, K\}$ that

$$c_k = \frac{(f_{\min\{k+1,K\}} - f_k)}{(\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_{\min\{k,K-1\}})} - \frac{(f_k - f_{\max\{k-1,0\}})}{(\mathfrak{x}_{\max\{k,1\}} - \mathfrak{x}_{\max\{k-1,0\}})} \tag{3.50}$$

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

$$\Phi_k = c_k \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k}). \tag{3.51}$$

Observe that Lemma 3.2.4 ensures that for all $k \in \{0, 1, \ldots, K\}$ it holds that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi_k) \in C(\mathbb{R}, \mathbb{R}), \qquad \mathcal{D}(\Phi_k) = (1, 1, 1) \in \mathbb{N}^3, \tag{3.52}$$

$$\text{and} \qquad \forall\, x \in \mathbb{R} \colon (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi_k))(x) = c_k \max\{x - \mathfrak{x}_k, 0\} \tag{3.53}$$

(cf. Definitions 1.2.4 and 1.3.4). This, Lemma 2.3.3, Lemma 2.4.11, and (3.49) establish that

$$\mathcal{D}(\mathbf{F}) = (1, K+1, 1) \in \mathbb{N}^3 \qquad \text{and} \qquad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}). \tag{3.54}$$

This proves item (i). Note that item (i) and (1.79) imply that

$$\mathcal{P}(\mathbf{F}) = 2(K+1) + (K+2) = 3K + 4. \tag{3.55}$$

This demonstrates item (iii). Observe that (3.50), (3.53), Lemma 2.3.3, and Lemma 2.4.11 show that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f_0 + \sum_{k=0}^{K} (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi_k))(x) = f_0 + \sum_{k=0}^{K} c_k \max\{x - \mathfrak{x}_k, 0\}. \tag{3.56}$$

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

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f_0 + 0 = f_0. \tag{3.57}$$

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

$$\sum_{n=0}^{k-1} c_n = \frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}. \tag{3.58}$$

We now prove (3.58) by induction on $k \in \{1, 2, \dots, K\}$. For the base case $k = 1$ observe that (3.50) establishes that

$$\sum_{n=0}^{0} c_n = c_0 = \frac{f_1 - f_0}{\mathfrak{x}_1 - \mathfrak{x}_0}. \tag{3.59}$$

This proves (3.58) in the base case $k = 1$. For the induction step note that (3.50) implies that for all $k \in \mathbb{N} \cap (1, \infty) \cap (0, K]$ with $\sum_{n=0}^{k-2} c_n = \frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}}$ it holds that

$$\sum_{n=0}^{k-1} c_n = c_{k-1} + \sum_{n=0}^{k-2} c_n = \frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} - \frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} + \frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} = \frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}. \tag{3.60}$$

Induction thus demonstrates (3.58). Furthermore, observe that (3.56), (3.58), and the fact that for all $k \in \{1, 2, \dots, K\}$ it holds that $\mathfrak{x}_{k-1} < \mathfrak{x}_k$ show that for all $k \in \{1, 2, \dots, K\}$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ it holds that

$$\begin{aligned}
(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) &= \sum_{n=0}^{K} c_n(\max\{x - \mathfrak{x}_n, 0\} - \max\{\mathfrak{x}_{k-1} - \mathfrak{x}_n, 0\}) \\
&= \sum_{n=0}^{k-1} c_n[(x - \mathfrak{x}_n) - (\mathfrak{x}_{k-1} - \mathfrak{x}_n)] = \sum_{n=0}^{k-1} c_n(x - \mathfrak{x}_{k-1}) \\
&= \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}\right)(x - \mathfrak{x}_{k-1}).
\end{aligned} \tag{3.61}$$

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

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f_{k-1} + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}\right)(x - \mathfrak{x}_{k-1}). \tag{3.62}$$

We now prove (3.62) by induction on $k \in \{1, 2, \dots, K\}$. For the base case $k = 1$ note that (3.57) and (3.61) ensure that for all $x \in [\mathfrak{x}_0, \mathfrak{x}_1]$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_0) + (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_0) = f_0 + \left(\frac{f_1 - f_0}{\mathfrak{x}_1 - \mathfrak{x}_0}\right)(x - \mathfrak{x}_0). \tag{3.63}$$

This establishes (3.62) in the base case $k = 1$. For the induction step observe that (3.61) proves that for all $k \in \mathbb{N} \cap (1, \infty) \cap [1, K]$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ with $\forall\, y \in [\mathfrak{x}_{k-2}, \mathfrak{x}_{k-1}]\colon (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(y) = f_{k-2} + \left(\frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}}\right)(y - \mathfrak{x}_{k-2})$ it holds that

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) &= (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(\mathfrak{x}_{k-1}) + (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(\mathfrak{x}_{k-1}) \\ &= f_{k-2} + \left(\frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}}\right)(\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}) + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}\right)(x - \mathfrak{x}_{k-1}) \\ &= f_{k-1} + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}\right)(x - \mathfrak{x}_{k-1}).\end{aligned} \tag{3.64}$$

Induction thus implies (3.62). Moreover, note that (3.50) and (3.58) demonstrate that

$$\sum_{n=0}^{K} c_n = c_K + \sum_{n=0}^{K-1} c_n = -\frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} + \frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} = 0. \tag{3.65}$$

The fact that for all $k \in \{0, 1, \ldots, K\}$ it holds that $\mathfrak{x}_k \leq \mathfrak{x}_K$ and (3.56) hence show that for all $x \in [\mathfrak{x}_K, \infty)$ it holds that

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(\mathfrak{x}_K) &= \left[\sum_{n=0}^{K} c_n(\max\{x - \mathfrak{x}_n, 0\} - \max\{\mathfrak{x}_K - \mathfrak{x}_n, 0\})\right] \\ &= \sum_{n=0}^{K} c_n[(x - \mathfrak{x}_n) - (\mathfrak{x}_K - \mathfrak{x}_n)] = \sum_{n=0}^{K} c_n(x - \mathfrak{x}_K) = 0.\end{aligned} \tag{3.66}$$

This and (3.62) ensure that for all $x \in [\mathfrak{x}_K, \infty)$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) = (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(\mathfrak{x}_K) = f_{K-1} + \left(\tfrac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}}\right)(\mathfrak{x}_K - \mathfrak{x}_{K-1}) = f_K. \tag{3.67}$$

Combining this, (3.57), (3.62), and (3.11) establishes item (ii). The proof of Proposition 3.2.5 is thus complete. □

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

$$\sup_{x \in [-2\pi, 2\pi]} \left|\cos(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x)\right| \leq \tfrac{1}{2} \tag{3.68}$$

(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 $\forall\, x, y, u, v \in \mathbb{R}\colon (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x, y, u, v) = \max\{x, y, u, v\}$ (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 $\forall\, x = (x_1, x_2, \ldots, x_{2^m}) \in \mathbb{R}\colon (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x) = \max\{x_1, x_2, \ldots, x_{2^m}\}$ (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 $K \in \mathbb{N}$, $L, a, \mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K \in \mathbb{R}$, $b \in (a, \infty)$ satisfy for all $k \in \{0, 1, \ldots, K\}$ that $\mathfrak{x}_k = a + \frac{k(b-a)}{K}$, let $f \colon [a, b] \to \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|, \tag{3.69}$$

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

$$\mathbf{F} = \mathbf{A}_{1, f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^{K} \left(\left(\tfrac{K(f(\mathfrak{x}_{\min\{k+1,K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1,0\}}))}{(b-a)} \right) \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1, -\mathfrak{x}_k}) \right) \right) \tag{3.70}$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.1). Then

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

(ii) it holds that $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) = \mathscr{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \ldots, f(\mathfrak{x}_K)}$,

(iii) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,

(iv) it holds that $\sup_{x \in [a,b]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq L(b - a)K^{-1}$, 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 $k \in \{0, 1, \ldots, K\}$ it holds that

$$\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_{\min\{k,K-1\}} = \mathfrak{x}_{\max\{k,1\}} - \mathfrak{x}_{\max\{k-1,0\}} = (b - a)K^{-1} \tag{3.71}$$

proves that for all $k \in \{0, 1, \ldots, K\}$ it holds that

$$\begin{aligned} &\frac{(f(\mathfrak{x}_{\min\{k+1,K\}}) - f(\mathfrak{x}_k))}{(\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_{\min\{k,K-1\}})} - \frac{(f(\mathfrak{x}_k) - f(\mathfrak{x}_{\max\{k-1,0\}}))}{(\mathfrak{x}_{\max\{k,1\}} - \mathfrak{x}_{\max\{k-1,0\}})} \\ &= \frac{K(f(\mathfrak{x}_{\min\{k+1,K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1,0\}}))}{(b - a)}. \end{aligned} \tag{3.72}$$

This and Proposition 3.2.5 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

$$|f(x) - f(y)| \leq L|x - y| \tag{3.73}$$

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

$$|f(x) - f(y)| \leq L|x - y|, \tag{3.74}$$

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

$$\mathfrak{x}_k - \mathfrak{x}_{k-1} = \frac{(b-a)}{K} \tag{3.75}$$

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

$$|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq L\left(\max_{k\in\{1,2,\ldots,K\}} |\mathfrak{x}_k - \mathfrak{x}_{k-1}|\right) = \frac{L(b-a)}{K}. \tag{3.76}$$

This proves item (iv). The proof of Proposition 3.3.1 is thus complete. □

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 $f \colon [a, b] \to \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|, \tag{3.77}$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\xi)} \bullet (0 \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\xi})) \tag{3.78}$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, and 3.2.1). Then

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

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

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

(iv) it holds that $\sup_{x\in[a,b]} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq L \max\{\xi - a, b - \xi\}$, 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 3.2.4 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

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) &= (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(0 \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\xi})))(x) + f(\xi) \\ &= 0\big((\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\xi}))(x)\big) + f(\xi) = f(\xi)\end{aligned} \tag{3.79}$$

(cf. Definitions 1.2.4 and 1.3.4). This establishes item (iii). Note that (3.79), the fact that $\xi \in [a,b]$, and the assumption that for all $x, y \in [a,b]$ it holds that

$$|f(x) - f(y)| \leq L|x-y| \tag{3.80}$$

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

$$|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| = |f(\xi) - f(x)| \leq L|x-\xi| \leq L\max\{\xi - a, b - \xi\}. \tag{3.81}$$

This proves item (iv). Observe that (1.79) and item (i) ensure that

$$\mathcal{P}(\mathbf{F}) = 1(1+1) + 1(1+1) = 4. \tag{3.82}$$

This establishes item (v). The proof of Lemma 3.3.2 is thus complete. □

Corollary 3.3.3 (Explicit ANN approximations with prescribed error tolerances)**.** *Let $\varepsilon \in (0,\infty)$, $L, a \in \mathbb{R}$, $b \in (a,\infty)$, $K \in \mathbb{N}_0 \cap \left[\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1\right)$, $\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K \in \mathbb{R}$ satisfy for all $k \in \{0,1,\ldots,K\}$ that $\mathfrak{x}_k = a + \frac{k(b-a)}{\max\{K,1\}}$, let $f\colon [a,b] \to \mathbb{R}$ satisfy for all $x, y \in [a,b]$ that*

$$|f(x) - f(y)| \leq L|x-y|, \tag{3.83}$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^{K} \left(\left(\tfrac{K(f(\mathfrak{x}_{\min\{k+1,K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1,0\}}))}{(b-a)} \right) \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k}) \right) \right) \tag{3.84}$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.1). Then

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

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

(iii) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(y)| \leq L|x-y|$,

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

$$\frac{L(b-a)}{\max\{K,1\}} \leq \varepsilon. \tag{3.85}$$

This, items (i), (iii), and (iv) in Proposition 3.3.1, and items (i), (ii), (iii), and (iv) in Lemma 3.3.2 prove 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

$$K \leq 1 + \frac{L(b-a)}{\varepsilon}, \tag{3.86}$$

demonstrate that

$$\mathcal{P}(\mathbf{F}) = 3K + 4 \leq \frac{3L(b-a)}{\varepsilon} + 7. \tag{3.87}$$

This establishes item (v). The proof of Corollary 3.3.3 is thus complete. □

3.3.2 Convergence rates for the approximation error

Definition 3.3.4 (Quasi vector norms). *We denote by* $\|\cdot\|_p \colon \left(\bigcup_{d=1}^{\infty} \mathbb{R}^d\right) \to \mathbb{R}$, $p \in (0, \infty]$, *the functions which satisfy for all* $p \in (0, \infty)$, $d \in \mathbb{N}$, $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ *that*

$$\|\theta\|_p = \left[\textstyle\sum_{i=1}^{d} |\theta_i|^p\right]^{1/p} \qquad \text{and} \qquad \|\theta\|_\infty = \max_{i \in \{1,2,\ldots,d\}} |\theta_i|. \tag{3.88}$$

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* $f \colon [a, b] \to \mathbb{R}$ *satisfy for all* $x, y \in [a, b]$ *that*

$$|f(x) - f(y)| \leq L|x - y|. \tag{3.89}$$

Then there exists $\mathbf{F} \in \mathbf{N}$ *such that*

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

(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* $|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,

(v) it holds that $\sup_{x \in [a,b]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$,

(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 $K \in \mathbb{N}_0 \cap \left[\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1\right)$, $\mathfrak{x}_0, \mathfrak{x}_1, \ldots, \mathfrak{x}_K \in [a, b]$, $c_0, c_1, \ldots, c_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \ldots, K\}$ that

$$\mathfrak{x}_k = a + \frac{k(b-a)}{\max\{K, 1\}} \qquad \text{and} \qquad c_k = \frac{K(f(\mathfrak{x}_{\min\{k+1,K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1,0\}}))}{(b-a)}, \tag{3.90}$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^{K} (c_k \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k})) \right) \tag{3.91}$$

(cf. Definitions 1.3.1, 2.1.1, 2.3.1, 2.3.4, 2.4.10, and 3.2.1). Note that Corollary 3.3.3 shows that

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

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

(III) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(y)| \leq L|x-y|$,

(IV) it holds that $\sup_{x\in[a,b]} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$, and

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

(cf. Definitions 1.2.4 and 1.3.4). This proves items (i), (iv), and (v). Observe that item (I) and the fact that

$$K \leq 1 + \frac{L(b-a)}{\varepsilon} \tag{3.92}$$

prove items (ii) and (iii). Note that item (iii) and items (I) and (V) ensure that

$$\mathcal{P}(\mathbf{F}) = 3K + 4 = 3(K+1) + 1 = 3(\mathbb{D}_1(\mathbf{F})) + 1 \leq \frac{3L(b-a)}{\varepsilon} + 7. \tag{3.93}$$

This establishes item (vi). Observe that Lemma 3.2.4 implies that for all $k \in \{0, 1, \ldots, K\}$ it holds that

$$c_k \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k}) = ((1, -\mathfrak{x}_k), (c_k, 0)). \tag{3.94}$$

Combining this with (2.155), (2.146), (2.137), and (2.2) demonstrates that

$$\mathbf{F} = \left(\left(\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} -\mathfrak{x}_0 \\ -\mathfrak{x}_1 \\ \vdots \\ -\mathfrak{x}_K \end{pmatrix} \right), \begin{pmatrix} c_0 & c_1 & \cdots & c_K \end{pmatrix}, f(\mathfrak{x}_0) \right) \in (\mathbb{R}^{(K+1)\times 1} \times \mathbb{R}^{K+1}) \times (\mathbb{R}^{1\times(K+1)} \times \mathbb{R}). \tag{3.95}$$

Lemma 1.3.9 therefore shows that

$$\|\mathcal{T}(\mathbf{F})\|_\infty = \max\{|\mathfrak{x}_0|, |\mathfrak{x}_1|, \ldots, |\mathfrak{x}_K|, |c_0|, |c_1|, \ldots, |c_K|, |f(\mathfrak{x}_0)|, 1\} \tag{3.96}$$

(cf. Definitions 1.3.6 and 3.3.4). Furthermore, note that (3.90), the assumption that for all $x, y \in [a, b]$ it holds that

$$|f(x) - f(y)| \leq L|x - y|, \tag{3.97}$$

and the fact that for all $k \in \mathbb{N} \cap (0, K + 1)$ it holds that

$$\mathfrak{x}_k - \mathfrak{x}_{k-1} = \frac{(b-a)}{\max\{K, 1\}} \tag{3.98}$$

prove that for all $k \in \{0, 1, \ldots, K\}$ it holds that

$$\begin{aligned} |c_k| &\leq \frac{K(|f(\mathfrak{x}_{\min\{k+1,K\}}) - f(\mathfrak{x}_k)| + |f(\mathfrak{x}_{\max\{k-1,0\}})) - f(\mathfrak{x}_k)|}{(b-a)} \\ &\leq \frac{KL(|\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_k| + |\mathfrak{x}_{\max\{k-1,0\}} - \mathfrak{x}_k|)}{(b-a)} \\ &\leq \frac{2KL(b-a)[\max\{K, 1\}]^{-1}}{(b-a)} \leq 2L. \end{aligned} \tag{3.99}$$

This and (3.96) establish item (vii). The proof of Corollary 3.3.5 is thus complete. □

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 $f\colon [a, b] \to \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|. \tag{3.100}$$

Then there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $\varepsilon \in (0, 1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \qquad \sup_{x\in[a,b]} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \qquad \mathcal{H}(\mathbf{F}) = 1, \tag{3.101}$$

$$\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, |a|, |b|, 2L, |f(a)|\}, \qquad \text{and} \qquad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-1} \tag{3.102}$$

(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

$$\mathfrak{C} = 3L(b - a) + 7. \tag{3.103}$$

Observe that the assumption that $a < b$ ensures that $L \geq 0$. Furthermore, note that (3.103) implies that for all $\varepsilon \in (0, 1]$ it holds that

$$3L(b-a)\varepsilon^{-1} + 7 \leq 3L(b-a)\varepsilon^{-1} + 7\varepsilon^{-1} = \mathfrak{C}\varepsilon^{-1}. \tag{3.104}$$

This and Corollary 3.3.5 demonstrate that for all $\varepsilon \in (0,1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R},\mathbb{R}), \qquad \sup_{x\in[a,b]}|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \qquad \mathcal{H}(\mathbf{F}) = 1, \tag{3.105}$$

$$\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1,|a|,|b|,2L,|f(a)|\}, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq 3L(b-a)\varepsilon^{-1} + 7 \leq \mathfrak{C}\varepsilon^{-1} \tag{3.106}$$

(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. □

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 $f\colon [a,b] \to \mathbb{R}$ satisfy for all $x, y \in [a,b]$ that*

$$|f(x) - f(y)| \leq L|x-y|. \tag{3.107}$$

Then there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $\varepsilon \in (0,1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R},\mathbb{R}), \quad \sup_{x\in[a,b]}|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-1} \tag{3.108}$$

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

Proof of Corollary 3.3.7. Observe that Corollary 3.3.6 proves (3.108). The proof of Corollary 3.3.7 is thus complete. □

Exercise 3.3.1. Let $f\colon [-2,3] \to \mathbb{R}$ satisfy for all $x \in [-2,3]$ that

$$f(x) = x^2 + 2\sin(x). \tag{3.109}$$

Prove or disprove the following statement: There exist $c \in \mathbb{R}$ and $\mathbf{F} = (\mathbf{F}_\varepsilon)_{\varepsilon\in(0,1]}\colon (0,1] \to \mathbf{N}$ such that for all $\varepsilon \in (0,1]$ it holds that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}_\varepsilon) \in C(\mathbb{R},\mathbb{R}), \quad \sup_{x\in[-2,3]}|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}_\varepsilon))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}_\varepsilon) \leq c\varepsilon^{-1} \tag{3.110}$$

(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

$$\sup_{x\in[0,10]} \left|\sqrt{x} - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x)\right| \leq \tfrac{1}{4} \tag{3.111}$$

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

Chapter 4

Multi-dimensional ANN approximation results

In this chapter we review basic deep ReLU ANN approximation results for possibly multi-dimensional target functions. We refer to the beginning of Chapter 3 for a small selection of ANN approximation results from the literature. The specific presentation of this chapter is strongly based on [25, Sections 2.2.6, 2.2.7, 2.2.8, and 3.1], [235, Sections 3 and 4.2], and [240, Section 3].

4.1 Approximations through supremal convolutions

Definition 4.1.1 (Metric). *We say that δ is a metric on E if and only if it holds that $\delta\colon E \times E \to [0,\infty)$ is a function from $E \times E$ to $[0,\infty)$ which satisfies that*

(i) it holds that

$$\{(x,y) \in E^2 \colon \delta(x,y) = 0\} = \textstyle\bigcup_{x \in E} \{(x,x)\} \tag{4.1}$$

(positive definiteness),

(ii) it holds for all $x, y \in E$ that

$$\delta(x,y) = \delta(y,x) \tag{4.2}$$

(symmetry), and

(iii) it holds for all $x, y, z \in E$ that

$$\delta(x,z) \leq \delta(x,y) + \delta(y,z) \tag{4.3}$$

(triangle inequality).

Definition 4.1.2 (Metric space)**.** *We say that $\mathcal{E}$ is a metric space if and only if there exist a set E and a metric δ on E such that*

$$\mathcal{E} = (E, \delta) \tag{4.4}$$

(cf. Definition 4.1.1).

Proposition 4.1.3 (Approximations through supremal convolutions)**.** *Let (E, δ) be a metric space, let $L \in [0, \infty)$, let $D \subseteq E$ and $\mathcal{M} \subseteq D$ satisfy $\mathcal{M} \neq \emptyset$, let $f \colon D \to \mathbb{R}$ satisfy for all $x \in D$, $y \in \mathcal{M}$ that $|f(x) - f(y)| \leq L\delta(x, y)$, and let $F \colon E \to \mathbb{R} \cup \{\infty\}$ satisfy for all $x \in E$ that*

$$F(x) = \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] \tag{4.5}$$

(cf. Definition 4.1.2). Then

(i) it holds for all $x \in \mathcal{M}$ that $F(x) = f(x)$,

(ii) it holds for all $x \in D$ that $F(x) \leq f(x)$,

(iii) it holds for all $x \in E$ that $F(x) < \infty$,

(iv) it holds for all $x, y \in E$ that $|F(x) - F(y)| \leq L\delta(x, y)$, and

(v) it holds for all $x \in D$ that

$$|F(x) - f(x)| \leq 2L \left[\inf_{y \in \mathcal{M}} \delta(x, y) \right]. \tag{4.6}$$

Proof of Proposition 4.1.3. First, note that the assumption that for all $x \in D$, $y \in \mathcal{M}$ it holds that $|f(x) - f(y)| \leq L\delta(x, y)$ ensures that for all $x \in D$, $y \in \mathcal{M}$ it holds that

$$f(y) + L\delta(x, y) \geq f(x) \geq f(y) - L\delta(x, y). \tag{4.7}$$

Hence, we obtain that for all $x \in D$ it holds that

$$f(x) \geq \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] = F(x). \tag{4.8}$$

This establishes item (ii). Moreover, note that (4.5) implies that for all $x \in \mathcal{M}$ it holds that

$$F(x) \geq f(x) - L\delta(x, x) = f(x). \tag{4.9}$$

This and (4.8) establish item (i). Observe that (4.7) (applied for every $y, z \in \mathcal{M}$ with $x \curvearrowleft y$, $y \curvearrowleft z$ in the notation of (4.7)) and the triangle inequality ensure that for all $x \in E$, $y, z \in \mathcal{M}$ it holds that

$$f(y) - L\delta(x,y) \leq f(z) + L\delta(y,z) - L\delta(x,y) \leq f(z) + L\delta(x,z). \tag{4.10}$$

Hence, we obtain that for all $x \in E$, $z \in \mathcal{M}$ it holds that

$$F(x) = \sup_{y \in \mathcal{M}} [f(y) - L\delta(x,y)] \leq f(z) + L\delta(x,z) < \infty. \tag{4.11}$$

This and the assumption that $\mathcal{M} \neq \emptyset$ prove item (iii). Note that item (iii), (4.5), and the triangle inequality show that for all $x, y \in E$ it holds that

$$\begin{aligned} F(x) - F(y) &= \left[\sup_{v \in \mathcal{M}} (f(v) - L\delta(x,v))\right] - \left[\sup_{w \in \mathcal{M}} (f(w) - L\delta(y,w))\right] \\ &= \sup_{v \in \mathcal{M}} \left[f(v) - L\delta(x,v) - \sup_{w \in \mathcal{M}} (f(w) - L\delta(y,w))\right] \\ &\leq \sup_{v \in \mathcal{M}} \left[f(v) - L\delta(x,v) - (f(v) - L\delta(y,v))\right] \\ &= \sup_{v \in \mathcal{M}} (L\delta(y,v) - L\delta(x,v)) \\ &\leq \sup_{v \in \mathcal{M}} (L\delta(y,x) + L\delta(x,v) - L\delta(x,v)) = L\delta(x,y). \end{aligned} \tag{4.12}$$

This and the fact that for all $x, y \in E$ it holds that $\delta(x,y) = \delta(y,x)$ establish item (iv). Observe that items (i) and (iv), the triangle inequality, and the assumption that $\forall\, x \in D, y \in \mathcal{M}\colon |f(x) - f(y)| \leq L\delta(x,y)$ ensure that for all $x \in D$ it holds that

$$\begin{aligned} |F(x) - f(x)| &= \inf_{y \in \mathcal{M}} |F(x) - F(y) + f(y) - f(x)| \\ &\leq \inf_{y \in \mathcal{M}} (|F(x) - F(y)| + |f(y) - f(x)|) \\ &\leq \inf_{y \in \mathcal{M}} (2L\delta(x,y)) = 2L\left[\inf_{y \in \mathcal{M}} \delta(x,y)\right]. \end{aligned} \tag{4.13}$$

This establishes item (v). The proof of Proposition 4.1.3 is thus complete. □

Corollary 4.1.4 (Approximations through supremum convolutions)**.** *Let (E, δ) be a metric space, let $L \in [0, \infty)$, let $\mathcal{M} \subseteq E$ satisfy $\mathcal{M} \neq \emptyset$, let $f\colon E \to \mathbb{R}$ satisfy for all $x \in E$, $y \in \mathcal{M}$ that $|f(x) - f(y)| \leq L\delta(x,y)$, and let $F\colon E \to \mathbb{R} \cup \{\infty\}$ satisfy for all $x \in E$ that*

$$F(x) = \sup_{y \in \mathcal{M}} [f(y) - L\delta(x,y)] \tag{4.14}$$

. Then

(i) *it holds for all* $x \in \mathcal{M}$ *that* $F(x) = f(x)$,

(ii) *it holds for all* $x \in E$ *that* $F(x) \leq f(x)$,

(iii) *it holds for all* $x, y \in E$ *that* $|F(x) - F(y)| \leq L\delta(x, y)$, *and*

(iv) *it holds for all* $x \in E$ *that*

$$|F(x) - f(x)| \leq 2L\left[\inf_{y \in \mathcal{M}} \delta(x, y)\right]. \tag{4.15}$$

Proof of Corollary 4.1.4. Note that Proposition 4.1.3 establishes items (i), (ii), (iii), and (iv). The proof of Corollary 4.1.4 is thus complete. □

Exercise 4.1.1. Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that $\mathcal{I}(\Phi) = 2$, $\mathcal{O}(\Phi) = 1$, $\mathcal{P}(\Phi) \leq 3\,000\,000\,000$, and

$$\sup_{x,y \in [0,2\pi]} |\sin(x)\sin(y) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x, y)| \leq \tfrac{1}{5}. \tag{4.16}$$

4.2 ANN representations

4.2.1 ANN representations for the 1-norm

Definition 4.2.1 (1-norm ANN representations). *We denote by* $(\mathbb{L}_d)_{d \in \mathbb{N}} \subseteq \mathbf{N}$ *the fully-connected feedforward ANNs which satisfy that*

(i) *it holds that*

$$\mathbb{L}_1 = \left(\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right), \left(\begin{pmatrix} 1 & 1 \end{pmatrix}, (0)\right)\right) \in (\mathbb{R}^{2\times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{1\times 2} \times \mathbb{R}^1) \tag{4.17}$$

and

(ii) *it holds for all* $d \in \{2, 3, 4, \dots\}$ *that* $\mathbb{L}_d = \mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)$

(cf. Definitions 1.3.1, 2.1.1, 2.2.1, and 2.4.1).

Proposition 4.2.2 (Properties of fully-connected feedforward 1-norm ANNs). *Let* $d \in \mathbb{N}$. *Then*

(i) *it holds that* $\mathcal{D}(\mathbb{L}_d) = (d, 2d, 1)$,

(ii) *it holds that* $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{L}_d) \in C(\mathbb{R}^d, \mathbb{R})$, *and*

(iii) it holds for all $x \in \mathbb{R}^d$ *that* $(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{L}_d))(x) = \|x\|_1$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 3.3.4, and 4.2.1).

Proof of Proposition 4.2.2. Note that the fact that $\mathcal{D}(\mathbb{L}_1) = (1, 2, 1)$ and Lemma 2.2.2 show that

$$\mathcal{D}(\mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \ldots, \mathbb{L}_1)) = (d, 2d, d) \tag{4.18}$$

(cf. Definitions 1.3.1, 2.2.1, and 4.2.1). Combining this, Proposition 2.1.2, and Lemma 2.3.2 ensures that

$$\mathcal{D}(\mathbb{L}_d) = \mathcal{D}\big(\mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \ldots, \mathbb{L}_1)\big) = (d, 2d, 1) \tag{4.19}$$

(cf. Definitions 2.1.1 and 2.4.1). This establishes item (i). Observe that (4.17) assures that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{L}_1))(x) = \mathfrak{r}(x) + \mathfrak{r}(-x) = \max\{x, 0\} + \max\{-x, 0\} = |x| = \|x\|_1 \tag{4.20}$$

(cf. Definitions 1.2.4, 1.3.4, and 3.3.4). Combining this and Proposition 2.2.3 shows that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ it holds that

$$\big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \ldots, \mathbb{L}_1))\big)(x) = (|x_1|, |x_2|, \ldots, |x_d|). \tag{4.21}$$

This and Lemma 2.4.2 demonstrate that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ it holds that

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{L}_d))(x) &= \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \ldots, \mathbb{L}_1))\big)(x) \\ &= \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{S}_{1,d})\big)(|x_1|, |x_2|, \ldots, |x_d|) = \sum_{k=1}^{d} |x_k| = \|x\|_1.\end{aligned} \tag{4.22}$$

This establishes items (ii) and (iii). The proof of Proposition 4.2.2 is thus complete. □

Lemma 4.2.3. *Let* $d \in \mathbb{N}$. *Then*

(i) it holds that $\mathcal{B}_{1,\mathbb{L}_d} = 0 \in \mathbb{R}^{2d}$,

(ii) it holds that $\mathcal{B}_{2,\mathbb{L}_d} = 0 \in \mathbb{R}$,

(iii) it holds that $\mathcal{W}_{1,\mathbb{L}_d} \in \{-1, 0, 1\}^{(2d)\times d}$,

(iv) it holds for all $x \in \mathbb{R}^d$ *that* $\|\mathcal{W}_{1,\mathbb{L}_d} x\|_\infty = \|x\|_\infty$, *and*

(v) it holds that $\mathcal{W}_{2,\mathbb{L}_d} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{1\times(2d)}$

(cf. Definitions 1.3.1, 3.3.4, and 4.2.1).

Proof of Lemma 4.2.3. Throughout this proof, assume without loss of generality that $d > 1$. Note that the fact that $\mathcal{B}_{1,\mathbb{L}_1} = 0 \in \mathbb{R}^2$, the fact that $\mathcal{B}_{2,\mathbb{L}_1} = 0 \in \mathbb{R}$, the fact that $\mathcal{B}_{1,\mathbb{S}_{1,d}} = 0 \in \mathbb{R}$, and the fact that $\mathbb{L}_d = \mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \ldots, \mathbb{L}_1)$ establish items (i) and (ii) (cf. Definitions 1.3.1, 2.1.1, 2.2.1, 2.4.1, and 4.2.1). In addition, observe that the fact that

$$\mathcal{W}_{1,\mathbb{L}_1} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad \text{and} \qquad \mathcal{W}_{1,\mathbb{L}_d} = \begin{pmatrix} \mathcal{W}_{1,\mathbb{L}_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{1,\mathbb{L}_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{1,\mathbb{L}_1} \end{pmatrix} \in \mathbb{R}^{(2d)\times d} \tag{4.23}$$

proves item (iii). Next note that (4.23) implies item (iv). Moreover, note that the fact that $\mathcal{W}_{2,\mathbb{L}_1} = (1\ 1)$ and the fact that $\mathbb{L}_d = \mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \ldots, \mathbb{L}_1)$ show that

$$\begin{aligned} \mathcal{W}_{2,\mathbb{L}_d} &= \mathcal{W}_{1,\mathbb{S}_{1,d}} \mathcal{W}_{2,\mathbf{P}_d(\mathbb{L}_1,\mathbb{L}_1,\ldots,\mathbb{L}_1)} \\ &= \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}}_{\in \mathbb{R}^{1\times d}} \underbrace{\begin{pmatrix} \mathcal{W}_{2,\mathbb{L}_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{2,\mathbb{L}_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{2,\mathbb{L}_1} \end{pmatrix}}_{\in \mathbb{R}^{d\times(2d)}} \\ &= \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{1\times(2d)}. \end{aligned} \tag{4.24}$$

This establishes item (v). The proof of Lemma 4.2.3 is thus complete. □

4.2.2 ANN representations for maxima

Lemma 4.2.4 (Unique existence of fully-connected feedforward maxima ANNs)**.** *There exist unique* $(\phi_d)_{d\in\mathbb{N}} \subseteq \mathbf{N}$ *which satisfy that*

(i) it holds for all $d \in \mathbb{N}$ *that* $\mathcal{I}(\phi_d) = d$,

(ii) it holds for all $d \in \mathbb{N}$ *that* $\mathcal{O}(\phi_d) = 1$,

(iii) it holds that $\phi_1 = \mathbf{A}_{1,0} \in \mathbb{R}^{1\times 1} \times \mathbb{R}^1$,

(iv) it holds that

$$\phi_2 = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right), \begin{pmatrix} (1 & 1 & -1), (0) \end{pmatrix} \right) \in (\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R}^1), \tag{4.25}$$

(v) it holds for all $d \in \{2, 3, 4, \ldots\}$ *that* $\phi_{2d} = \phi_d \bullet \big(\mathbf{P}_d(\phi_2, \phi_2, \ldots, \phi_2)\big)$, *and*

(vi) it holds for all $d \in \{2, 3, 4, \ldots\}$ *that* $\phi_{2d-1} = \phi_d \bullet \big(\mathbf{P}_d(\phi_2, \phi_2, \ldots, \phi_2, \mathfrak{I}_1)\big)$

(cf. Definitions 1.3.1, 2.1.1, 2.2.1, 2.2.6, and 2.3.1).

Proof of Lemma 4.2.4. Throughout this proof, let $\psi \in \mathbf{N}$ satisfy

$$\psi = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\right), \big(\begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, \big(0\big)\big)\right) \in (\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R}^1) \tag{4.26}$$

(cf. Definition 1.3.1). Note that (4.26) and Lemma 2.2.7 show that

$$\mathcal{I}(\psi) = 2, \qquad \mathcal{O}(\psi) = \mathcal{I}(\mathfrak{I}_1) = \mathcal{O}(\mathfrak{I}_1) = 1, \qquad \text{and} \qquad \mathcal{L}(\psi) = \mathcal{L}(\mathfrak{I}_1) = 2. \tag{4.27}$$

Lemma 2.2.2 and Lemma 2.2.7 hence establish that for all $d \in \mathbb{N} \cap (1, \infty)$ it holds that

$$\mathcal{I}(\mathbf{P}_d(\psi, \psi, \ldots, \psi)) = 2d, \qquad \mathcal{O}(\mathbf{P}_d(\psi, \psi, \ldots, \psi)) = d, \tag{4.28}$$

$$\mathcal{I}(\mathbf{P}_d(\psi, \psi, \ldots, \psi, \mathfrak{I}_1)) = 2d - 1, \qquad \text{and} \qquad \mathcal{O}(\mathbf{P}_d(\psi, \psi, \ldots, \psi, \mathfrak{I}_1)) = d \tag{4.29}$$

(cf. Definitions 2.2.1 and 2.2.6). Combining (4.27), Proposition 2.1.2, and induction therefore ensures that there exists unique $\phi_d \in \mathbf{N}$, $d \in \mathbb{N}$, which satisfy for all $d \in \mathbb{N}$ that $\mathcal{I}(\phi_d) = d$, $\mathcal{O}(\phi_d) = 1$, and

$$\phi_d = \begin{cases} \mathbf{A}_{1,0} & : d = 1 \\ \psi & : d = 2 \\ \phi_{d/2} \bullet \big(\mathbf{P}_{d/2}(\psi, \psi, \ldots, \psi)\big) & : d \in \{4, 6, 8, \ldots\} \\ \phi_{(d+1)/2} \bullet \big(\mathbf{P}_{(d+1)/2}(\psi, \psi, \ldots, \psi, \mathfrak{I}_1)\big) & : d \in \{3, 5, 7, \ldots\}. \end{cases} \tag{4.30}$$

The proof of Lemma 4.2.4 is thus complete. □

Definition 4.2.5 (Maxima ANN representations)**.** *We denote by* $(\mathbb{M}_d)_{d\in\mathbb{N}} \subseteq \mathbf{N}$ *the fully-connected feedforward ANNs which satisfy that*

(i) it holds for all $d \in \mathbb{N}$ *that* $\mathcal{I}(\mathbb{M}_d) = d$,

(ii) it holds for all $d \in \mathbb{N}$ *that* $\mathcal{O}(\mathbb{M}_d) = 1$,

(iii) it holds that $\mathbb{M}_1 = \mathbf{A}_{1,0} \in \mathbb{R}^{1\times 1} \times \mathbb{R}^1$,

(iv) it holds that

$$\mathbb{M}_2 = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\right), \big(\begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, \big(0\big)\big)\right) \in (\mathbb{R}^{3\times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1\times 3} \times \mathbb{R}^1), \tag{4.31}$$

(v) it holds for all $d \in \{2, 3, 4, \ldots\}$ *that* $\mathbb{M}_{2d} = \mathbb{M}_d \bullet \big(\mathbf{P}_d(\mathbb{M}_2, \mathbb{M}_2, \ldots, \mathbb{M}_2)\big)$*, and*

(vi) it holds for all $d \in \{2, 3, 4, \ldots\}$ *that* $\mathbb{M}_{2d-1} = \mathbb{M}_d \bullet \big(\mathbf{P}_d(\mathbb{M}_2, \mathbb{M}_2, \ldots, \mathbb{M}_2, \mathfrak{I}_1)\big)$

(cf. Definitions 1.3.1, 2.1.1, 2.2.1, 2.2.6, and 2.3.1 and Lemma 4.2.4).

Definition 4.2.6 (Floor and ceiling of real numbers)**.** *We denote by* $\lceil \cdot \rceil \colon \mathbb{R} \to \mathbb{Z}$ *and* $\lfloor \cdot \rfloor \colon \mathbb{R} \to \mathbb{Z}$ *the functions which satisfy for all* $x \in \mathbb{R}$ *that*

$$\lceil x \rceil = \min(\mathbb{Z} \cap [x, \infty)) \qquad \text{and} \qquad \lfloor x \rfloor = \max(\mathbb{Z} \cap (-\infty, x]). \tag{4.32}$$

Exercise 4.2.1. Prove or disprove the following statement: For all $n \in \{3, 5, 7, \ldots\}$ it holds that $\lceil \log_2(n+1) \rceil = \lceil \log_2(n) \rceil$.

Proposition 4.2.7 (Properties of fully-connected feedforward maxima ANNs)**.** *Let* $d \in \mathbb{N}$*. Then*

(i) it holds that $\mathcal{H}(\mathbb{M}_d) = \lceil \log_2(d) \rceil$*,*

(ii) it holds for all $i \in \mathbb{N}$ *that* $\mathbb{D}_i(\mathbb{M}_d) \leq 3\lceil \frac{d}{2^i} \rceil$*,*

(iii) it holds that $\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_d) \in C(\mathbb{R}^d, \mathbb{R})$*, and*

(iv) it holds for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ *that* $(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_d))(x) = \max\{x_1, x_2, \ldots, x_d\}$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 4.2.5, and 4.2.6).

Proof of Proposition 4.2.7. Throughout this proof, assume without loss of generality that $d > 1$. Note that (4.31) ensures that

$$\mathcal{H}(\mathbb{M}_2) = 1 \tag{4.33}$$

(cf. Definitions 1.3.1 and 4.2.5). This and (2.44) demonstrate that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that

$$\mathcal{H}(\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \ldots, \mathbb{M}_2)) = \mathcal{H}(\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \ldots, \mathbb{M}_2, \mathfrak{I}_1)) = \mathcal{H}(\mathbb{M}_2) = 1 \tag{4.34}$$

(cf. Definitions 2.2.1 and 2.2.6). Combining this with Proposition 2.1.2 establishes that for all $\mathfrak{d} \in \{3, 4, 5, \ldots\}$ it holds that

$$\mathcal{H}(\mathbb{M}_{\mathfrak{d}}) = \mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) + 1 \tag{4.35}$$

(cf. Definition 4.2.6). This assures that for all $\mathfrak{d} \in \{4, 6, 8, \ldots\}$ with $\mathcal{H}(\mathbb{M}_{\mathfrak{d}/2}) = \lceil \log_2(\mathfrak{d}/2) \rceil$ it holds that

$$\begin{aligned} \mathcal{H}(\mathbb{M}_{\mathfrak{d}}) &= \mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) + 1 = \mathcal{H}(\mathbb{M}_{\mathfrak{d}/2}) + 1 \\ &= \lceil \log_2(\mathfrak{d}/2) \rceil + 1 = \lceil \log_2(\mathfrak{d}) - 1 \rceil + 1 = \lceil \log_2(\mathfrak{d}) \rceil. \end{aligned} \tag{4.36}$$

Furthermore, observe that (4.35) and the fact that for all $\mathfrak{d} \in \{3, 5, 7, \ldots\}$ it holds that $\lceil \log_2(\mathfrak{d}+1) \rceil = \lceil \log_2(\mathfrak{d}) \rceil$ imply that for all $\mathfrak{d} \in \{3, 5, 7, \ldots\}$ with $\mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) = \lceil \log_2(\lceil \mathfrak{d}/2 \rceil) \rceil$ it holds that

$$\begin{aligned} \mathcal{H}(\mathbb{M}_{\mathfrak{d}}) &= \mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) + 1 = \big\lceil \log_2(\lceil \mathfrak{d}/2 \rceil) \big\rceil + 1 = \big\lceil \log_2((\mathfrak{d}+1)/2) \big\rceil + 1 \\ &= \lceil \log_2(\mathfrak{d}+1) - 1 \rceil + 1 = \lceil \log_2(\mathfrak{d}+1) \rceil = \lceil \log_2(\mathfrak{d}) \rceil. \end{aligned} \tag{4.37}$$

Combining this and (4.36) demonstrates that for all $\mathfrak{d} \in \{3, 4, 5, \ldots\}$ with $\forall\, k \in \{2, 3, \ldots, \mathfrak{d}-1\}\colon \mathcal{H}(\mathbb{M}_k) = \lceil \log_2(k) \rceil$ it holds that

$$\mathcal{H}(\mathbb{M}_{\mathfrak{d}}) = \lceil \log_2(\mathfrak{d}) \rceil. \tag{4.38}$$

The fact that $\mathcal{H}(\mathbb{M}_2) = 1$ and induction hence establish item (i). Note that the fact that $\mathcal{D}(\mathbb{M}_2) = (2, 3, 1)$ assure that for all $i \in \mathbb{N}$ it holds that

$$\mathbb{D}_i(\mathbb{M}_2) \leq 3 = 3\big\lceil \tfrac{2}{2^i} \big\rceil. \tag{4.39}$$

Moreover, observe that Proposition 2.1.2 and Lemma 2.2.2 imply that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$, $i \in \mathbb{N}$ it holds that

$$\mathbb{D}_i(\mathbb{M}_{2\mathfrak{d}}) = \mathbb{D}_i\big(\mathbb{M}_{\mathfrak{d}} \bullet \big(\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \ldots, \mathbb{M}_2)\big)\big) = \begin{cases} 3\mathfrak{d} & : i = 1 \\ \mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}}) & : i \geq 2 \end{cases} \tag{4.40}$$

and

$$\mathbb{D}_i(\mathbb{M}_{2\mathfrak{d}-1}) = \mathbb{D}_i\big(\mathbb{M}_{\mathfrak{d}} \bullet \big(\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \ldots, \mathbb{M}_2, \mathfrak{I}_1)\big)\big) = \begin{cases} 3\mathfrak{d} - 1 & : i = 1 \\ \mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}}) & : i \geq 2. \end{cases} \tag{4.41}$$

This and (4.38) assure that for all $\mathfrak{d} \in \{2, 4, 6, \ldots\}$ it holds that

$$\mathbb{D}_1(\mathbb{M}_{\mathfrak{d}}) = 3(\tfrac{\mathfrak{d}}{2}) = 3\big\lceil \tfrac{\mathfrak{d}}{2} \big\rceil. \tag{4.42}$$

In addition, note that (4.41) shows that for all $\mathfrak{d} \in \{3, 5, 7, \ldots\}$ it holds that

$$\mathbb{D}_1(\mathbb{M}_{\mathfrak{d}}) = 3\big\lceil \tfrac{\mathfrak{d}}{2} \big\rceil - 1 \leq 3\big\lceil \tfrac{\mathfrak{d}}{2} \big\rceil. \tag{4.43}$$

This and (4.42) show that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that

$$\mathbb{D}_1(\mathbb{M}_{\mathfrak{d}}) \leq 3\big\lceil \tfrac{\mathfrak{d}}{2} \big\rceil. \tag{4.44}$$

Next observe that (4.40) demonstrates that for all $\mathfrak{d} \in \{4, 6, 8, \ldots\}$, $i \in \{2, 3, 4, \ldots\}$ with $\mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}/2}) \leq 3\lceil(\mathfrak{d}/2)\frac{1}{2^{i-1}}\rceil$ it holds that

$$\mathbb{D}_i(\mathbb{M}_{\mathfrak{d}}) = \mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}/2}) \leq 3\lceil(\mathfrak{d}/2)\tfrac{1}{2^{i-1}}\rceil = 3\lceil\tfrac{\mathfrak{d}}{2^i}\rceil. \tag{4.45}$$

Furthermore, note that (4.41) and the fact that for all $\mathfrak{d} \in \{3, 5, 7, \ldots\}$, $i \in \mathbb{N}$ it holds that $\lceil\frac{\mathfrak{d}+1}{2^i}\rceil = \lceil\frac{\mathfrak{d}}{2^i}\rceil$ establish that for all $\mathfrak{d} \in \{3, 5, 7, \ldots\}$, $i \in \{2, 3, 4, \ldots\}$ with $\mathbb{D}_{i-1}(\mathbb{M}_{\lceil\mathfrak{d}/2\rceil}) \leq 3\lceil\lceil\mathfrak{d}/2\rceil\frac{1}{2^{i-1}}\rceil$ it holds that

$$\mathbb{D}_i(\mathbb{M}_{\mathfrak{d}}) = \mathbb{D}_{i-1}(\mathbb{M}_{\lceil\mathfrak{d}/2\rceil}) \leq 3\lceil\lceil\mathfrak{d}/2\rceil\tfrac{1}{2^{i-1}}\rceil = 3\lceil\tfrac{\mathfrak{d}+1}{2^i}\rceil = 3\lceil\tfrac{\mathfrak{d}}{2^i}\rceil. \tag{4.46}$$

This, (4.44), and (4.45) ensure that for all $\mathfrak{d} \in \{3, 4, 5, \ldots\}$, $i \in \mathbb{N}$ with $\forall\, k \in \{2, 3, \ldots, \mathfrak{d} - 1\}, j \in \mathbb{N}\colon \mathbb{D}_j(\mathbb{M}_k) \leq 3\lceil\frac{k}{2^j}\rceil$ it holds that

$$\mathbb{D}_i(\mathbb{M}_{\mathfrak{d}}) \leq 3\lceil\tfrac{\mathfrak{d}}{2^i}\rceil. \tag{4.47}$$

Combining this and (4.39) with induction establishes item (ii). Observe that (4.31) ensures that for all $x = (x_1, x_2) \in \mathbb{R}^2$ it holds that

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_2))(x) &= \max\{x_1 - x_2, 0\} + \max\{x_2, 0\} - \max\{-x_2, 0\}\\ &= \max\{x_1 - x_2, 0\} + x_2 = \max\{x_1, x_2\}\end{aligned} \tag{4.48}$$

(cf. Definitions 1.2.4, 1.3.4, and 2.1.1). Proposition 2.2.3, Proposition 2.1.2, Lemma 2.2.7, and induction hence imply that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$, $x = (x_1, x_2, \ldots, x_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_{\mathfrak{d}}) \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}) \qquad \text{and} \qquad \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_{\mathfrak{d}})\big)(x) = \max\{x_1, x_2, \ldots, x_{\mathfrak{d}}\}. \tag{4.49}$$

This establishes items (iii) and (iv). The proof of Proposition 4.2.7 is thus complete. □

Lemma 4.2.8. *Let $d \in \mathbb{N}$, $i \in \{1, 2, \ldots, \mathcal{L}(\mathbb{M}_d)\}$ (cf. Definitions 1.3.1 and 4.2.5). Then*

(i) it holds that $\mathcal{B}_{i,\mathbb{M}_d} = 0 \in \mathbb{R}^{\mathbb{D}_i(\mathbb{M}_d)}$,

(ii) it holds that $\mathcal{W}_{i,\mathbb{M}_d} \in \{-1, 0, 1\}^{\mathbb{D}_i(\mathbb{M}_d) \times \mathbb{D}_{i-1}(\mathbb{M}_d)}$, and

(iii) it holds for all $x \in \mathbb{R}^d$ that $\|\mathcal{W}_{1,\mathbb{M}_d} x\|_\infty \leq 2\|x\|_\infty$

(cf. Definition 3.3.4).

Proof of Lemma 4.2.8. Throughout this proof, assume without loss of generality that $d > 2$ (cf. items (iii) and (iv) in Definition 4.2.5) and let $A_1 \in \mathbb{R}^{3\times 2}$, $A_2 \in \mathbb{R}^{1\times 3}$, $C_1 \in \mathbb{R}^{2\times 1}$,

$C_2 \in \mathbb{R}^{1\times 2}$ satisfy

$$A_1 = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, \qquad C_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \text{and} \qquad C_2 = \begin{pmatrix} 1 & -1 \end{pmatrix}. \tag{4.50}$$

Note that items (iv), (v), and (vi) in Definition 4.2.5 assure that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that

$$\mathcal{W}_{1,\mathbb{M}_{2\mathfrak{d}-1}} = \underbrace{\begin{pmatrix} A_1 & 0 & \cdots & 0 & 0 \\ 0 & A_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_1 & 0 \\ 0 & 0 & \cdots & 0 & C_1 \end{pmatrix}}_{\in \mathbb{R}^{(3\mathfrak{d}-1)\times(2\mathfrak{d}-1)}}, \qquad \mathcal{W}_{1,\mathbb{M}_{2\mathfrak{d}}} = \underbrace{\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_1 \end{pmatrix}}_{\in \mathbb{R}^{(3\mathfrak{d})\times(2\mathfrak{d})}}, \tag{4.51}$$

$$\mathcal{B}_{1,\mathbb{M}_{2\mathfrak{d}-1}} = 0 \in \mathbb{R}^{3\mathfrak{d}-1}, \qquad \text{and} \qquad \mathcal{B}_{1,\mathbb{M}_{2\mathfrak{d}}} = 0 \in \mathbb{R}^{3\mathfrak{d}}.$$

This and (4.50) proves item (iii). Furthermore, note that (4.51) and item (iv) in Definition 4.2.5 imply that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that $\mathcal{B}_{1,\mathbb{M}_\mathfrak{d}} = 0$. Items (iv), (v), and (vi) in Definition 4.2.5 hence ensure that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that

$$\mathcal{W}_{2,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{W}_{1,\mathbb{M}_\mathfrak{d}} \underbrace{\begin{pmatrix} A_2 & 0 & \cdots & 0 & 0 \\ 0 & A_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_2 & 0 \\ 0 & 0 & \cdots & 0 & C_2 \end{pmatrix}}_{\in \mathbb{R}^{\mathfrak{d}\times(3\mathfrak{d}-1)}}, \qquad \mathcal{W}_{2,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{W}_{1,\mathbb{M}_\mathfrak{d}} \underbrace{\begin{pmatrix} A_2 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_2 \end{pmatrix}}_{\in \mathbb{R}^{\mathfrak{d}\times(3\mathfrak{d})}},$$

$$\mathcal{B}_{2,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{B}_{1,\mathbb{M}_\mathfrak{d}} = 0, \qquad \text{and} \qquad \mathcal{B}_{2,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{B}_{1,\mathbb{M}_\mathfrak{d}} = 0. \tag{4.52}$$

Combining this and item (iv) in Definition 4.2.5 shows that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that $\mathcal{B}_{2,\mathbb{M}_\mathfrak{d}} = 0$. Moreover, note that (2.2) demonstrates that for all $\mathfrak{d} \in \{2, 3, 4, \ldots, \}$, $i \in \{3, 4, \ldots, \mathcal{L}(\mathbb{M}_\mathfrak{d}) + 1\}$ it holds that

$$\mathcal{W}_{i,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{W}_{i,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{W}_{i-1,\mathbb{M}_\mathfrak{d}} \qquad \text{and} \qquad \mathcal{B}_{i,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{B}_{i,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{B}_{i-1,\mathbb{M}_\mathfrak{d}}. \tag{4.53}$$

This, (4.50), (4.51), (4.52), the fact that for all $\mathfrak{d} \in \{2, 3, 4, \ldots\}$ it holds that $\mathcal{B}_{2,\mathbb{M}_\mathfrak{d}} = 0$, and induction establish items (i) and (ii). The proof of Lemma 4.2.8 is thus complete. □

4.2.3 ANN representations for maximum convolutions

Exercise 4.2.2. Prove or disprove the following statement: It holds for all $d \in \mathbb{N}$, $x \in \mathbb{R}^d$ that

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_d \bullet \mathbf{P}_d(\mathbb{L}_1, \ldots, \mathbb{L}_1))(x) = \|x\|_\infty \tag{4.54}$$

(cf. Definitions 1.2.4, 1.3.4, 2.1.1, 2.2.1, 3.3.4, 4.2.1, and 4.2.5).

Lemma 4.2.9. *Let $d, K \in \mathbb{N}$, $L \in [0, \infty)$, $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in \mathbb{R}^d$, $\mathfrak{y} = (\mathfrak{y}_1, \ldots, \mathfrak{y}_K) \in \mathbb{R}^K$, $\Phi \in \mathbf{N}$ satisfy*

$$\Phi = \mathbb{M}_K \bullet \mathbf{A}_{-L\mathrm{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K\big(\mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_2}, \ldots, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_K}\big) \bullet \mathbb{T}_{d,K} \tag{4.55}$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, 2.2.1, 2.3.1, 2.4.6, 4.2.1, and 4.2.5). Then

(i) it holds that $\mathcal{I}(\Phi) = d$,

(ii) it holds that $\mathcal{O}(\Phi) = 1$,

(iii) it holds that $\mathcal{H}(\Phi) = \lceil \log_2(K) \rceil + 1$,

(iv) it holds that $\mathbb{D}_1(\Phi) = 2dK$,

(v) it holds for all $i \in \{2, 3, 4, \ldots\}$ that $\mathbb{D}_i(\Phi) \leq 3\lceil \frac{K}{2^{i-1}} \rceil$,

(vi) it holds that $\|\mathcal{T}(\Phi)\|_\infty \leq \max\{1, L, \max_{k \in \{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty, 2\|\mathfrak{y}\|_\infty\}$, and

(vii) it holds for all $x \in \mathbb{R}^d$ that $(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x) = \max_{k \in \{1,2,\ldots,K\}}(\mathfrak{y}_k - L\|x - \mathfrak{x}_k\|_1)$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, 3.3.4, and 4.2.6).

Proof of Lemma 4.2.9. Throughout this proof, let $\Psi_k \in \mathbf{N}$, $k \in \{1, 2, \ldots, K\}$, satisfy for all $k \in \{1, 2, \ldots, K\}$ that $\Psi_k = \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_k}$, let $\Xi \in \mathbf{N}$ satisfy

$$\Xi = \mathbf{A}_{-L\mathrm{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K\big(\Psi_1, \Psi_2, \ldots, \Psi_K\big) \bullet \mathbb{T}_{d,K}, \tag{4.56}$$

and let $|||\cdot|||\colon \bigcup_{m,n \in \mathbb{N}} \mathbb{R}^{m \times n} \to [0, \infty)$ satisfy for all $m, n \in \mathbb{N}$, $M = (M_{i,j})_{i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}} \in \mathbb{R}^{m \times n}$ that $|||M||| = \max_{i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}} |M_{i,j}|$. Observe that (4.55) and Proposition 2.1.2 ensure that $\mathcal{O}(\Phi) = \mathcal{O}(\mathbb{M}_K) = 1$ and $\mathcal{I}(\Phi) = \mathcal{I}(\mathbb{T}_{d,K}) = d$. This proves items (i) and (ii). Moreover, observe that the fact that for all $m, n \in \mathbb{N}$, $\mathfrak{W} \in \mathbb{R}^{m \times n}$, $\mathfrak{B} \in \mathbb{R}^m$ it holds that $\mathcal{H}(\mathbf{A}_{\mathfrak{W}, \mathfrak{B}}) = 0 = \mathcal{H}(\mathbb{T}_{d,K})$, the fact that $\mathcal{H}(\mathbb{L}_d) = 1$, and Proposition 2.1.2 assure that

$$\mathcal{H}(\Xi) = \mathcal{H}(\mathbf{A}_{-L\mathrm{I}_K, \mathfrak{y}}) + \mathcal{H}(\mathbf{P}_K(\Psi_1, \Psi_2, \ldots, \Psi_K)) + \mathcal{H}(\mathbb{T}_{d,K}) = \mathcal{H}(\Psi_1) = \mathcal{H}(\mathbb{L}_d) = 1. \tag{4.57}$$

Proposition 2.1.2 and Proposition 4.2.7 hence ensure that

$$\mathcal{H}(\Phi) = \mathcal{H}(\mathbb{M}_K \bullet \Xi) = \mathcal{H}(\mathbb{M}_K) + \mathcal{H}(\Xi) = \lceil \log_2(K) \rceil + 1 \tag{4.58}$$

(cf. Definition 4.2.6). This establishes item (iii). Next observe that the fact that $\mathcal{H}(\Xi) = 1$, Proposition 2.1.2, and Proposition 4.2.7 assure that for all $i \in \{2, 3, 4, \ldots\}$ it holds that

$$\mathbb{D}_i(\Phi) = \mathbb{D}_{i-1}(\mathbb{M}_K) \leq 3\left\lceil \tfrac{K}{2^{i-1}} \right\rceil. \tag{4.59}$$

This proves item (v). Furthermore, note that Proposition 2.1.2, Proposition 2.2.4, and Proposition 4.2.2 assure that

$$\mathbb{D}_1(\Phi) = \mathbb{D}_1(\Xi) = \mathbb{D}_1(\mathbf{P}_K(\Psi_1, \Psi_2, \ldots, \Psi_K)) = \sum_{i=1}^{K} \mathbb{D}_1(\Psi_i) = \sum_{i=1}^{K} \mathbb{D}_1(\mathbb{L}_d) = 2dK. \tag{4.60}$$

This establishes item (iv). Moreover, observe that (2.2) and Lemma 4.2.8 imply that

$$\begin{aligned} \Phi = \big((\mathcal{W}_{1,\Xi}, \mathcal{B}_{1,\Xi}), (\mathcal{W}_{1,\mathbb{M}_K}\mathcal{W}_{2,\Xi}, \mathcal{W}_{1,\mathbb{M}_K}\mathcal{B}_{2,\Xi}), \\ (\mathcal{W}_{2,\mathbb{M}_K}, 0), \ldots, (\mathcal{W}_{\mathcal{L}(\mathbb{M}_K),\mathbb{M}_K}, 0)\big). \end{aligned} \tag{4.61}$$

Next note that the fact that for all $k \in \{1, 2, \ldots, K\}$ it holds that $\mathcal{W}_{1,\Psi_k} = \mathcal{W}_{1,\mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_k}}\mathcal{W}_{1,\mathbb{L}_d} = \mathcal{W}_{1,\mathbb{L}_d}$ assures that

$$\begin{aligned} \mathcal{W}_{1,\Xi} = \mathcal{W}_{1,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)}\mathcal{W}_{1,\mathbb{T}_{d,K}} &= \begin{pmatrix} \mathcal{W}_{1,\Psi_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{1,\Psi_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{1,\Psi_K} \end{pmatrix} \begin{pmatrix} \mathrm{I}_d \\ \mathrm{I}_d \\ \vdots \\ \mathrm{I}_d \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{W}_{1,\Psi_1} \\ \mathcal{W}_{1,\Psi_2} \\ \vdots \\ \mathcal{W}_{1,\Psi_K} \end{pmatrix} = \begin{pmatrix} \mathcal{W}_{1,\mathbb{L}_d} \\ \mathcal{W}_{1,\mathbb{L}_d} \\ \vdots \\ \mathcal{W}_{1,\mathbb{L}_d} \end{pmatrix}. \end{aligned} \tag{4.62}$$

Lemma 4.2.3 hence demonstrates that $|||\mathcal{W}_{1,\Xi}||| = 1$. In addition, note that (2.2) implies that

$$\mathcal{B}_{1,\Xi} = \mathcal{W}_{1,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)}\mathcal{B}_{1,\mathbb{T}_{d,K}} + \mathcal{B}_{1,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)} = \mathcal{B}_{1,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)} = \begin{pmatrix} \mathcal{B}_{1,\Psi_1} \\ \mathcal{B}_{1,\Psi_2} \\ \vdots \\ \mathcal{B}_{1,\Psi_K} \end{pmatrix}. \tag{4.63}$$

Furthermore, observe that Lemma 4.2.3 implies that for all $k \in \{1, 2, \ldots, K\}$ it holds that

$$\mathcal{B}_{1,\Psi_k} = \mathcal{W}_{1,\mathbb{L}_d}\mathcal{B}_{1,\mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_k}} + \mathcal{B}_{1,\mathbb{L}_d} = -\mathcal{W}_{1,\mathbb{L}_d}\mathfrak{x}_k. \tag{4.64}$$

This, (4.63), and Lemma 4.2.3 show that

$$\|\mathcal{B}_{1,\Xi}\|_\infty = \max_{k\in\{1,2,\ldots,K\}} \|\mathcal{B}_{1,\Psi_k}\|_\infty = \max_{k\in\{1,2,\ldots,K\}} \|\mathcal{W}_{1,\mathbb{L}_d}\mathfrak{x}_k\|_\infty = \max_{k\in\{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty \tag{4.65}$$

(cf. Definition 3.3.4). Combining this, (4.61), Lemma 4.2.8, and the fact that $|||\mathcal{W}_{1,\Xi}||| = 1$ shows that

$$\begin{aligned}\|\mathcal{T}(\Phi)\|_\infty &= \max\{|||\mathcal{W}_{1,\Xi}|||, \|\mathcal{B}_{1,\Xi}\|_\infty, |||\mathcal{W}_{1,\mathbb{M}_K}\mathcal{W}_{2,\Xi}|||, \|\mathcal{W}_{1,\mathbb{M}_K}\mathcal{B}_{2,\Xi}\|_\infty, 1\}\\ &= \max\{1, \max_{k\in\{1,2,\ldots,K\}}\|\mathfrak{x}_k\|_\infty, |||\mathcal{W}_{1,\mathbb{M}_K}\mathcal{W}_{2,\Xi}|||, \|\mathcal{W}_{1,\mathbb{M}_K}\mathcal{B}_{2,\Xi}\|_\infty\}\end{aligned} \tag{4.66}$$

(cf. Definition 1.3.6). Next note that Lemma 4.2.3 ensures that for all $k \in \{1, 2, \ldots, K\}$ it holds that $\mathcal{B}_{2,\Psi_k} = \mathcal{B}_{2,\mathbb{L}_d} = 0$. Hence, we obtain that $\mathcal{B}_{2,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)} = 0$. This implies that

$$\mathcal{B}_{2,\Xi} = \mathcal{W}_{1,\mathbf{A}_{-L\mathrm{I}_K,\mathfrak{y}}}\mathcal{B}_{2,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)} + \mathcal{B}_{1,\mathbf{A}_{-L\mathrm{I}_K,\mathfrak{y}}} = \mathcal{B}_{1,\mathbf{A}_{-L\mathrm{I}_K,\mathfrak{y}}} = \mathfrak{y}. \tag{4.67}$$

In addition, observe that the fact that for all $k \in \{1, 2, \ldots, K\}$ it holds that $\mathcal{W}_{2,\Psi_k} = \mathcal{W}_{2,\mathbb{L}_d}$ assures that

$$\begin{aligned}\mathcal{W}_{2,\Xi} &= \mathcal{W}_{1,\mathbf{A}_{-L\mathrm{I}_K,\mathfrak{y}}}\mathcal{W}_{2,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)} = -L\mathcal{W}_{2,\mathbf{P}_K(\Psi_1,\Psi_2,\ldots,\Psi_K)}\\ &= -L\begin{pmatrix}\mathcal{W}_{2,\Psi_1} & 0 & \cdots & 0\\ 0 & \mathcal{W}_{2,\Psi_2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \mathcal{W}_{2,\Psi_K}\end{pmatrix} = \begin{pmatrix}-L\mathcal{W}_{2,\mathbb{L}_d} & 0 & \cdots & 0\\ 0 & -L\mathcal{W}_{2,\mathbb{L}_d} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & -L\mathcal{W}_{2,\mathbb{L}_d}\end{pmatrix}.\end{aligned} \tag{4.68}$$

Item (v) in Lemma 4.2.3 and Lemma 4.2.8 hence imply that

$$|||\mathcal{W}_{1,\mathbb{M}_K}\mathcal{W}_{2,\Xi}||| = L|||\mathcal{W}_{1,\mathbb{M}_K}||| \leq L. \tag{4.69}$$

Moreover, observe that (4.67) and Lemma 4.2.8 show that

$$\|\mathcal{W}_{1,\mathbb{M}_K}\mathcal{B}_{2,\Xi}\|_\infty \leq 2\|\mathcal{B}_{2,\Xi}\|_\infty = 2\|\mathfrak{y}\|_\infty. \tag{4.70}$$

Combining this with (4.66) and (4.69) establishes item (vi). Next observe that Proposition 4.2.2 and Lemma 2.3.3 show that for all $x \in \mathbb{R}^d$, $k \in \{1, 2, \ldots, K\}$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Psi_k))(x) = \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{L}_d) \circ \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_k})\big)(x) = \|x - \mathfrak{x}_k\|_1. \tag{4.71}$$

This, Proposition 2.2.3, and Proposition 2.1.2 imply that for all $x \in \mathbb{R}^d$ it holds that

$$\big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{P}_K(\Psi_1, \Psi_2, \ldots, \Psi_K) \bullet \mathbb{T}_{d,K})\big)(x) = \big(\|x - \mathfrak{x}_1\|_1, \|x - \mathfrak{x}_2\|_1, \ldots, \|x - \mathfrak{x}_K\|_1\big). \tag{4.72}$$

(cf. Definitions 1.2.4 and 1.3.4). Combining this and Lemma 2.3.3 establishes that for all $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Xi))(x) &= \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{A}_{-L\mathrm{I}_K,\mathfrak{y}}) \circ \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{P}_K(\Psi_1, \Psi_2, \ldots, \Psi_K) \bullet \mathbb{T}_{d,K})\big)(x)\\ &= \big(\mathfrak{y}_1 - L\|x - \mathfrak{x}_1\|_1, \mathfrak{y}_2 - L\|x - \mathfrak{x}_2\|_1, \ldots, \mathfrak{y}_K - L\|x - \mathfrak{x}_K\|_1\big).\end{aligned} \tag{4.73}$$

Proposition 2.1.2 and Proposition 4.2.7 hence demonstrate that for all $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned}(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x) &= \big(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_K) \circ \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Xi)\big)(x)\\ &= (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbb{M}_K))\big(\mathfrak{y}_1 - L\|x-\mathfrak{x}_1\|_1, \mathfrak{y}_2 - L\|x-\mathfrak{x}_2\|_1, \ldots, \mathfrak{y}_K - L\|x-\mathfrak{x}_K\|_1\big)\\ &= \max_{k\in\{1,2,\ldots,K\}}(\mathfrak{y}_k - L\|x-\mathfrak{x}_k\|_1).\end{aligned} \tag{4.74}$$

This establishes item (vii). The proof of Lemma 4.2.9 is thus complete. □

4.3 ANN approximations results for multi-dimensional functions

4.3.1 Constructive ANN approximation results

Proposition 4.3.1. *Let $d, K \in \mathbb{N}$, $L \in [0,\infty)$, let $E \subseteq \mathbb{R}^d$ be a set, let $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in E$, let $f\colon E \to \mathbb{R}$ satisfy for all $x, y \in E$ that $|f(x) - f(y)| \leq L\|x-y\|_1$, and let $\mathfrak{y} \in \mathbb{R}^K$, $\Phi \in \mathbf{N}$ satisfy $\mathfrak{y} = (f(\mathfrak{x}_1), f(\mathfrak{x}_2), \ldots, f(\mathfrak{x}_K))$ and*

$$\Phi = \mathbb{M}_K \bullet \mathbf{A}_{-L\,\mathrm{I}_K,\mathfrak{y}} \bullet \mathbf{P}_K\big(\mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_2}, \ldots, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_K}\big) \bullet \mathbb{T}_{d,K} \tag{4.75}$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, 2.2.1, 2.3.1, 2.4.6, 3.3.4, 4.2.1, and 4.2.5). Then

$$\sup_{x\in E}|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x) - f(x)| \leq 2L\big[\sup_{x\in E}\big(\min_{k\in\{1,2,\ldots,K\}}\|x-\mathfrak{x}_k\|_1\big)\big] \tag{4.76}$$

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Proposition 4.3.1. Throughout this proof, let $F\colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$F(x) = \max_{k\in\{1,2,\ldots,K\}}(f(\mathfrak{x}_k) - L\|x-\mathfrak{x}_k\|_1). \tag{4.77}$$

Observe that Corollary 4.1.4, (4.77), and the assumption that for all $x, y \in E$ it holds that $|f(x) - f(y)| \leq L\|x-y\|_1$ establish that

$$\sup_{x\in E}|F(x) - f(x)| \leq 2L\big[\sup_{x\in E}\big(\min_{k\in\{1,2,\ldots,K\}}\|x-\mathfrak{x}_k\|_1\big)\big]. \tag{4.78}$$

Moreover, note that Lemma 4.2.9 ensures that for all $x \in E$ it holds that $F(x) = (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x)$. Combining this and (4.78) establishes (4.76). The proof of Proposition 4.3.1 is thus complete. □

Exercise 4.3.1. Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that $\mathcal{I}(\Phi) = 2$, $\mathcal{O}(\Phi) = 1$, $\mathcal{P}(\Phi) < 20$, and

$$\sup_{v=(x,y)\in[0,2]^2}\big|x^2 + y^2 - 2x - 2y + 2 - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(v)\big| \leq \tfrac{3}{8}. \tag{4.79}$$

4.3.2 Covering number estimates

Definition 4.3.2 (Covering numbers)**.** *Let (E,δ) be a metric space and let $r \in [0,\infty]$. Then we denote by $\mathcal{C}^{(E,\delta),r} \in \mathbb{N}_0 \cup \{\infty\}$ (we denote by $\mathcal{C}^{E,r} \in \mathbb{N}_0 \cup \{\infty\}$) the extended real number given by*

$$\mathcal{C}^{(E,\delta),r} = \min\left(\left\{n \in \mathbb{N}_0 \colon \left[\exists\, A \subseteq E \colon \begin{pmatrix} (|A| \leq n) \wedge (\forall\, x \in E \colon \\ \exists\, a \in A \colon \delta(a,x) \leq r) \end{pmatrix}\right]\right\} \cup \{\infty\}\right) \tag{4.80}$$

and we call $\mathcal{C}^{(E,\delta),r}$ the r-covering number of (E,δ) (we call $\mathcal{C}^{E,r}$ the r-covering number of E).

Lemma 4.3.3. *Let (E,δ) be a metric space and let $r \in [0,\infty]$. Then*

$$\mathcal{C}^{(E,\delta),r} = \begin{cases} 0 & : X = \emptyset \\ \inf\left(\left\{n \in \mathbb{N} \colon \left(\exists\, x_1, x_2, \ldots, x_n \in E \colon \right.\right.\right. & \\ \left.\left.\left. E \subseteq \left[\bigcup_{m=1}^{n} \{v \in E \colon d(x_m, v) \leq r\}\right]\right)\right\} \cup \{\infty\}\right) & : X \neq \emptyset \end{cases} \tag{4.81}$$

(cf. Definition 4.3.2).

Proof of Lemma 4.3.3. Throughout this proof, assume without loss of generality that $E \neq \emptyset$. Observe that Lemma 12.2.4 establishes (4.81). The proof of Lemma 4.3.3 is thus complete. □

Exercise 4.3.2. Prove or disprove the following statement: For every metric space (X,d), every $Y \subseteq X$, and every $r \in [0,\infty]$ it holds that $\mathcal{C}^{(Y,d|_{Y\times Y}),r} \leq \mathcal{C}^{(X,d),r}$.

Exercise 4.3.3. Prove or disprove the following statement: For every metric space (E,δ) it holds that $\mathcal{C}^{(E,\delta),\infty} = 1$.

Exercise 4.3.4. Prove or disprove the following statement: For every metric space (E,δ) and every $r \in [0,\infty)$ with $\mathcal{C}^{(E,\delta),r} < \infty$ it holds that E is bounded. (*Note:* A metric space (E,δ) is bounded if and only if there exists $r \in [0,\infty)$ such that it holds for all $x,y \in E$ that $\delta(x,y) \leq r$.)

Exercise 4.3.5. Prove or disprove the following statement: For every bounded metric space (E,δ) and every $r \in [0,\infty]$ it holds that $\mathcal{C}^{(E,\delta),r} < \infty$.

Lemma 4.3.4. *Let $d \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a,\infty)$, $r \in (0,\infty)$ and for every $p \in [1,\infty)$ let $\delta_p \colon ([a,b]^d) \times ([a,b]^d) \to [0,\infty)$ satisfy for all $x,y \in [a,b]^d$ that $\delta_p(x,y) = \|x-y\|_p$ (cf.*

Definition 3.3.4). Then it holds for all $p \in [1,\infty)$ that

$$\mathcal{C}^{([a,b]^d,\delta_p),r} \leq \left(\left\lceil \tfrac{d^{1/p}(b-a)}{2r} \right\rceil\right)^d \leq \begin{cases} 1 & : r \geq {}^{d(b-a)}\!/_{2} \\ \left(\frac{d(b-a)}{r}\right)^d & : r < {}^{d(b-a)}\!/_{2}. \end{cases} \tag{4.82}$$

(cf. Definitions 4.2.6 and 4.3.2).

Proof of Lemma 4.3.4. Throughout this proof, let $(\mathfrak{N}_p)_{p\in[1,\infty)} \subseteq \mathbb{N}$ satisfy for all $p \in [1,\infty)$ that

$$\mathfrak{N}_p = \left\lceil \tfrac{d^{1/p}(b-a)}{2r} \right\rceil, \tag{4.83}$$

for every $N \in \mathbb{N}$, $i \in \{1,2,\dots,N\}$ let $g_{N,i} \in [a,b]$ be given by

$$g_{N,i} = a + {}^{(i-1/2)(b-a)}\!/_{N} \tag{4.84}$$

and for every $p \in [1,\infty)$ let $A_p \subseteq [a,b]^d$ be given by

$$A_p = \{g_{\mathfrak{N}_p,1}, g_{\mathfrak{N}_p,2}, \dots, g_{\mathfrak{N}_p,\mathfrak{N}_p}\}^d \tag{4.85}$$

(cf. Definition 4.2.6). Observe that it holds for all $N \in \mathbb{N}$, $i \in \{1,2,\dots,N\}$, $x \in [a + {}^{(i-1)(b-a)}\!/_{N}, g_{N,i}]$ that

$$|x - g_{N,i}| = a + \tfrac{(i-1/2)(b-a)}{N} - x \leq a + \tfrac{(i-1/2)(b-a)}{N} - \left(a + \tfrac{(i-1)(b-a)}{N}\right) = \tfrac{b-a}{2N}. \tag{4.86}$$

In addition, note that it holds for all $N \in \mathbb{N}$, $i \in \{1,2,\dots,N\}$, $x \in [g_{N,i}, a + {}^{i(b-a)}\!/_{N}]$ that

$$|x - g_{N,i}| = x - \left(a + \tfrac{(i-1/2)(b-a)}{N}\right) \leq a + \tfrac{i(b-a)}{N} - \left(a + \tfrac{(i-1/2)(b-a)}{N}\right) = \tfrac{b-a}{2N}. \tag{4.87}$$

Combining this with (4.86) implies for all $N \in \mathbb{N}$, $i \in \{1,2,\dots,N\}$, $x \in [a + {}^{(i-1)(b-a)}\!/_{N}, a + {}^{i(b-a)}\!/_{N}]$ that $|x - g_{N,i}| \leq {}^{(b-a)}\!/_{(2N)}$. This proves that for every $N \in \mathbb{N}$, $x \in [a,b]$ there exists $y \in \{g_{N,1}, g_{N,2}, \dots, g_{N,N}\}$ such that

$$|x - y| \leq \tfrac{b-a}{2N}. \tag{4.88}$$

This establishes that for every $p \in [1,\infty)$, $x = (x_1,\dots,x_d) \in [a,b]^d$ there exists $y = (y_1,\dots,y_d) \in A_p$ such that

$$\delta_p(x,y) = \|x-y\|_p = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{d} \tfrac{(b-a)^p}{(2\mathfrak{N}_p)^p}\right)^{1/p} = \tfrac{d^{1/p}(b-a)}{2\mathfrak{N}_p} \leq \tfrac{d^{1/p}(b-a)2r}{2d^{1/p}(b-a)} = r. \tag{4.89}$$

Combining this with (4.80), (4.85), (4.83), and the fact that $\forall\, x \in [0,\infty)\colon \lceil x \rceil \leq \mathbb{1}_{(0,1]}(x) + 2x\mathbb{1}_{(1,\infty)}(x) = \mathbb{1}_{(0,r]}(rx) + 2x\mathbb{1}_{(r,\infty)}(rx)$ yields that for all $p \in [1,\infty)$ it holds that

$$\begin{aligned} \mathcal{C}^{([a,b]^d,\delta_p),r} &\leq |A_p| = (\mathfrak{N}_p)^d = \left(\left\lceil \tfrac{d^{1/p}(b-a)}{2r} \right\rceil\right)^d \leq \left(\left\lceil \tfrac{d(b-a)}{2r} \right\rceil\right)^d \\ &\leq \left(\mathbb{1}_{(0,r]}\left(\tfrac{d(b-a)}{2}\right) + \tfrac{2d(b-a)}{2r}\mathbb{1}_{(r,\infty)}\left(\tfrac{d(b-a)}{2}\right)\right)^d \\ &= \mathbb{1}_{(0,r]}\left(\tfrac{d(b-a)}{2}\right) + \left(\tfrac{d(b-a)}{r}\right)^d \mathbb{1}_{(r,\infty)}\left(\tfrac{d(b-a)}{2}\right) \end{aligned} \tag{4.90}$$

(cf. Definition 4.3.2). The proof of Lemma 4.3.4 is thus complete. □

4.3.3 Convergence rates for the approximation error

Lemma 4.3.5. *Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, let $f\colon [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that $|f(x) - f(y)| \leq L\|x - y\|_1$, and let $\mathbf{F} = \mathbf{A}_{0, f((a+b)/2, (a+b)/2, \ldots, (a+b)/2)} \in \mathbb{R}^{1\times d} \times \mathbb{R}^1$ (cf. Definitions 2.3.1 and 3.3.4). Then*

(i) it holds that $\mathcal{I}(\mathbf{F}) = d$,

(ii) it holds that $\mathcal{O}(\mathbf{F}) = 1$,

(iii) it holds that $\mathcal{H}(\mathbf{F}) = 0$,

(iv) it holds that $\mathcal{P}(\mathbf{F}) = d + 1$,

(v) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \sup_{x\in[a,b]^d} |f(x)|$, and

(vi) it holds that $\sup_{x\in[a,b]^d} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \frac{dL(b-a)}{2}$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Lemma 4.3.5. Note that the assumption that for all $x, y \in [a,b]^d$ it holds that $|f(x) - f(y)| \leq L\|x - y\|_1$ assures that $L \geq 0$. Next observe that Lemma 2.3.2 assures that for all $x \in \mathbb{R}^d$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) = f\big((a+b)/2, (a+b)/2, \ldots, (a+b)/2\big). \tag{4.91}$$

The fact that for all $x \in [a,b]$ it holds that $|x - (a+b)/2| \leq (b-a)/2$ and the assumption that for all $x, y \in [a,b]^d$ it holds that $|f(x) - f(y)| \leq L\|x - y\|_1$ hence ensure that for all $x = (x_1, \ldots, x_d) \in [a,b]^d$ it holds that

$$\begin{aligned} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| &= |f\big((a+b)/2, (a+b)/2, \ldots, (a+b)/2\big) - f(x)| \\ &\leq L\big\|\big((a+b)/2, (a+b)/2, \ldots, (a+b)/2\big) - x\big\|_1 \\ &= L\sum_{i=1}^{d} |(a+b)/2 - x_i| \leq \sum_{i=1}^{d} \tfrac{L(b-a)}{2} = \tfrac{dL(b-a)}{2}. \end{aligned} \tag{4.92}$$

This and the fact that $\|\mathcal{T}(\mathbf{F})\|_\infty = |f((a+b)/2, (a+b)/2, \ldots, (a+b)/2)| \leq \sup_{x\in[a,b]^d} |f(x)|$ complete The proof of Lemma 4.3.5 is thus complete. □

Proposition 4.3.6. *Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $r \in (0, d/4)$, let $f\colon [a,b]^d \to \mathbb{R}$ and $\delta\colon [a,b]^d \times [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that $|f(x) - f(y)| \leq L\|x - y\|_1$ and $\delta(x,y) = \|x - y\|_1$, and let $K \in \mathbb{N}$, $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in [a,b]^d$, $\mathfrak{y} \in \mathbb{R}^K$, $\mathbf{F} \in \mathbf{N}$ satisfy $K = \mathcal{C}^{([a,b]^d, \delta), (b-a)r}$, $\sup_{x\in[a,b]^d} \big[\min_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\big] \leq (b-a)r$, $\mathfrak{y} = (f(\mathfrak{x}_1), f(\mathfrak{x}_2), \ldots, f(\mathfrak{x}_K))$,*

and

$$\mathbf{F} = \mathbb{M}_K \bullet \mathbf{A}_{-L\,\mathrm{I}_K,\mathfrak{y}} \bullet \mathbf{P}_K\big(\mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_2}, \ldots, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d,-\mathfrak{x}_K}\big) \bullet \mathbb{T}_{d,K} \tag{4.93}$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, 2.2.1, 2.3.1, 2.4.6, 3.3.4, 4.2.1, 4.2.5, and 4.3.2). Then

(i) it holds that $\mathcal{I}(\mathbf{F}) = d$,

(ii) it holds that $\mathcal{O}(\mathbf{F}) = 1$,

(iii) it holds that $\mathcal{H}(\mathbf{F}) \leq \lceil d \log_2\big(\frac{3d}{4r}\big) \rceil + 1$,

(iv) it holds that $\mathbb{D}_1(\mathbf{F}) \leq 2d\big(\frac{3d}{4r}\big)^d$,

(v) it holds for all $i \in \{2, 3, 4, \ldots\}$ *that* $\mathbb{D}_i(\mathbf{F}) \leq 3\big\lceil \big(\frac{3d}{4r}\big)^d \frac{1}{2^{i-1}} \big\rceil$,

(vi) it holds that $\mathcal{P}(\mathbf{F}) \leq 35\big(\frac{3d}{4r}\big)^{2d} d^2$,

(vii) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d} |f(x)|]\}$, *and*

(viii) it holds that $\sup_{x\in[a,b]^d} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq 2L(b-a)r$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 4.2.6).

Proof of Proposition 4.3.6. Note that the assumption that for all $x, y \in [a,b]^d$ it holds that $|f(x) - f(y)| \leq L\|x - y\|_1$ assures that $L \geq 0$. Next observe that (4.93), Lemma 4.2.9, and Proposition 4.3.1 demonstrate that

(I) it holds that $\mathcal{I}(\mathbf{F}) = d$,

(II) it holds that $\mathcal{O}(\mathbf{F}) = 1$,

(III) it holds that $\mathcal{H}(\mathbf{F}) = \lceil \log_2(K) \rceil + 1$,

(IV) it holds that $\mathbb{D}_1(\mathbf{F}) = 2dK$,

(V) it holds for all $i \in \{2, 3, 4, \ldots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3\big\lceil \frac{K}{2^{i-1}} \big\rceil$,

(VI) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, \max_{k\in\{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty, 2[\max_{k\in\{1,2,\ldots,K\}} |f(\mathfrak{x}_k)|]\}$, and

(VII) it holds that $\sup_{x\in[a,b]^d} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq 2L\big[\sup_{x\in[a,b]^d} \big(\min_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\big)\big]$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 4.2.6). Note that items (I) and (II) establish items (i) and (ii). Next observe that Lemma 4.3.4 and the fact that $\frac{d}{2r} \geq 2$ prove that

$$K = \mathcal{C}^{([a,b]^d,\delta),(b-a)r} \leq \left(\left\lceil \tfrac{d(b-a)}{2(b-a)r} \right\rceil\right)^d = \left(\left\lceil \tfrac{d}{2r} \right\rceil\right)^d \leq \left(\tfrac{3}{2}\left(\tfrac{d}{2r}\right)\right)^d = \left(\tfrac{3d}{4r}\right)^d. \tag{4.94}$$

Combining this with item (III) assures that

$$\mathcal{H}(\mathbf{F}) = \lceil \log_2(K) \rceil + 1 \leq \left\lceil \log_2\left(\left(\tfrac{3d}{4r}\right)^d\right) \right\rceil + 1 = \lceil d \log_2\left(\tfrac{3d}{4r}\right) \rceil + 1. \tag{4.95}$$

This establishes item (iii). Moreover, note that (4.94) and item (IV) imply that

$$\mathbb{D}_1(\mathbf{F}) = 2dK \leq 2d\left(\tfrac{3d}{4r}\right)^d. \tag{4.96}$$

This establishes item (iv). In addition, observe that item (V) and (4.94) establish item (v). Next note that item (III) ensures that for all $i \in \mathbb{N} \cap (1, \mathcal{H}(\mathbf{F})]$ it holds that

$$\tfrac{K}{2^{i-1}} \geq \tfrac{K}{2^{\mathcal{H}(\mathbf{F})-1}} = \tfrac{K}{2^{\lceil \log_2(K) \rceil}} \geq \tfrac{K}{2^{\log_2(K)+1}} = \tfrac{K}{2K} = \tfrac{1}{2}. \tag{4.97}$$

Item (V) and (4.94) hence show that for all $i \in \mathbb{N} \cap (1, \mathcal{H}(\mathbf{F})]$ it holds that

$$\mathbb{D}_i(\mathbf{F}) \leq 3\left\lceil \tfrac{K}{2^{i-1}} \right\rceil \leq \tfrac{3K}{2^{i-2}} \leq \left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{i-2}}. \tag{4.98}$$

Furthermore, note that the fact that for all $x \in [a,b]^d$ it holds that $\|x\|_\infty \leq \max\{|a|, |b|\}$ and item (VI) imply that

$$\begin{aligned} \|\mathcal{T}(\mathbf{F})\|_\infty &\leq \max\{1, L, \max_{k\in\{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty, 2[\max_{k\in\{1,2,\ldots,K\}} |f(\mathfrak{x}_k)|]\} \\ &\leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d} |f(x)|]\}. \end{aligned} \tag{4.99}$$

This establishes item (vii). Moreover, observe that the assumption that

$$\sup_{x\in[a,b]^d} \left[\min_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\right] \leq (b-a)r \tag{4.100}$$

and item (VII) demonstrate that

$$\sup_{x\in[a,b]^d} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq 2L\left[\sup_{x\in[a,b]^d} \left(\min_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\right)\right] \leq 2L(b-a)r. \tag{4.101}$$

This establishes item (viii). It thus remains to prove item (vi). For this note that items (I) and (II), (4.96), and (4.98) show that

$$\begin{aligned} \mathcal{P}(\mathbf{F}) &= \sum_{i=1}^{\mathcal{L}(\mathbf{F})} \mathbb{D}_i(\mathbf{F})(\mathbb{D}_{i-1}(\mathbf{F}) + 1) \\ &\leq 2d\left(\tfrac{3d}{4r}\right)^d (d+1) + \left(\tfrac{3d}{4r}\right)^d 3\left(2d\left(\tfrac{3d}{4r}\right)^d + 1\right) \\ &\quad + \left[\sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{i-2}} \left(\left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{i-3}} + 1\right)\right] + \left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{\mathcal{L}(\mathbf{F})-3}} + 1. \end{aligned} \tag{4.102}$$

Next note that the fact that $\frac{3d}{4r} \geq 3$ ensures that

$$\begin{aligned} &2d\left(\tfrac{3d}{4r}\right)^d(d+1) + \left(\tfrac{3d}{4r}\right)^d 3\left(2d\left(\tfrac{3d}{4r}\right)^d + 1\right) + \left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{\mathcal{L}(\mathbf{F})-3}} + 1 \\ &\leq \left(\tfrac{3d}{4r}\right)^{2d}\left(2d(d+1) + 3(2d+1) + \tfrac{3}{2^{1-3}} + 1\right) \\ &\leq \left(\tfrac{3d}{4r}\right)^{2d} d^2(4+9+12+1) = 26\left(\tfrac{3d}{4r}\right)^{2d} d^2. \end{aligned} \tag{4.103}$$

Moreover, observe that the fact that $\frac{3d}{4r} \geq 3$ implies that

$$\begin{aligned} \sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{i-2}}\left(\left(\tfrac{3d}{4r}\right)^d \tfrac{3}{2^{i-3}} + 1\right) &\leq \left(\tfrac{3d}{4r}\right)^{2d} \sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \tfrac{3}{2^{i-2}}\left(\tfrac{3}{2^{i-3}} + 1\right) \\ &= \left(\tfrac{3d}{4r}\right)^{2d} \sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \left[\tfrac{9}{2^{2i-5}} + \tfrac{3}{2^{i-2}}\right] \\ &= \left(\tfrac{3d}{4r}\right)^{2d} \sum_{i=0}^{\mathcal{L}(\mathbf{F})-4} \left[\tfrac{9}{2}(4^{-i}) + \tfrac{3}{2}(2^{-i})\right] \\ &\leq \left(\tfrac{3d}{4r}\right)^{2d}\left(\tfrac{9}{2}\left(\tfrac{1}{1-4^{-1}}\right) + \tfrac{3}{2}\left(\tfrac{1}{1-2^{-1}}\right)\right) = 9\left(\tfrac{3d}{4r}\right)^{2d}. \end{aligned} \tag{4.104}$$

Combining this, (4.102), and (4.103) demonstrates that

$$\mathcal{P}(\mathbf{F}) \leq 26\left(\tfrac{3d}{4r}\right)^{2d} d^2 + 9\left(\tfrac{3d}{4r}\right)^{2d} \leq 35\left(\tfrac{3d}{4r}\right)^{2d} d^2. \tag{4.105}$$

This establishes item (vi). The proof of Proposition 4.3.6 is thus complete. □

Proposition 4.3.7. *Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $r \in (0, \infty)$ and let $f\colon [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that $|f(x) - f(y)| \leq L\|x - y\|_1$ (cf. Definition 3.3.4). Then there exists $\mathbf{F} \in \mathbf{N}$ such that*

(i) it holds that $\mathcal{I}(\mathbf{F}) = d$,

(ii) it holds that $\mathcal{O}(\mathbf{F}) = 1$,

(iii) it holds that $\mathcal{H}(\mathbf{F}) \leq \left(\left\lceil d\log_2\left(\frac{3d}{4r}\right)\right\rceil + 1\right)\mathbb{1}_{(0,d/4)}(r)$,

(iv) it holds that $\mathbb{D}_1(\mathbf{F}) \leq 2d\left(\frac{3d}{4r}\right)^d \mathbb{1}_{(0,d/4)}(r) + \mathbb{1}_{[d/4,\infty)}(r)$,

(v) it holds for all $i \in \{2, 3, 4, \ldots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3\left\lceil \left(\frac{3d}{4r}\right)^d \frac{1}{2^{i-1}}\right\rceil$,

(vi) it holds that $\mathcal{P}(\mathbf{F}) \leq 35\left(\frac{3d}{4r}\right)^{2d} d^2 \mathbb{1}_{(0,d/4)}(r) + (d+1)\mathbb{1}_{[d/4,\infty)}(r)$,

(vii) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d}|f(x)|]\}$, and

(viii) it holds that $\sup_{x\in[a,b]^d}|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq 2L(b-a)r$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 4.2.6).

Proof of Proposition 4.3.7. Throughout this proof, assume without loss of generality that $r < {}^{d}\!/_{4}$ (cf. Lemma 4.3.5), let $\delta\colon [a,b]^d \times [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that

$$\delta(x,y) = \|x-y\|_1, \tag{4.106}$$

and let $K \in \mathbb{N} \cup \{\infty\}$ satisfy

$$K = \mathcal{C}^{([a,b]^d,\delta),(b-a)r}. \tag{4.107}$$

Note that Lemma 4.3.4 assures that $K < \infty$. This and (4.80) ensure that there exist $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in [a,b]^d$ such that

$$\sup_{x\in[a,b]^d}\big[\min_{k\in\{1,2,\ldots,K\}}\delta(x,\mathfrak{x}_k)\big] \leq (b-a)r. \tag{4.108}$$

Combining this with Proposition 4.3.6 establishes items (i), (ii), (iii), (iv), (v), (vi), (vii), and (viii). The proof of Proposition 4.3.7 is thus complete. □

Proposition 4.3.8 (Implicit multi-dimensional ANN approximations with prescribed error tolerances and explicit parameter bounds)**.** *Let* $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in [a,\infty)$, $\varepsilon \in (0,1]$ *and let* $f\colon [a,b]^d \to \mathbb{R}$ *satisfy for all* $x, y \in [a,b]^d$ *that*

$$|f(x) - f(y)| \leq L\|x-y\|_1 \tag{4.109}$$

(cf. Definition 3.3.4). Then there exists $\mathbf{F} \in \mathbf{N}$ *such that*

(i) it holds that $\mathcal{I}(\mathbf{F}) = d$,

(ii) it holds that $\mathcal{O}(\mathbf{F}) = 1$,

(iii) it holds that $\mathcal{H}(\mathbf{F}) \leq d\big(\log_2\big(\max\big\{\frac{3dL(b-a)}{2}, 1\big\}\big) + \log_2(\varepsilon^{-1})\big) + 2$,

(iv) it holds that $\mathbb{D}_1(\mathbf{F}) \leq \varepsilon^{-d} d(3d\max\{L(b-a),1\})^d$,

(v) it holds for all $i \in \{2,3,4,\ldots\}$ *that* $\mathbb{D}_i(\mathbf{F}) \leq \varepsilon^{-d}3\big(\frac{(3dL(b-a))^d}{2^i} + 1\big)$,

(vi) it holds that $\mathcal{P}(\mathbf{F}) \leq \varepsilon^{-2d}9\big(3d\max\{L(b-a),1\}\big)^{2d}d^2$,

(vii) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d}|f(x)|]\}$, *and*

(viii) it holds that $\sup_{x\in[a,b]^d}|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Proposition 4.3.8. Throughout this proof, assume without loss of generality that

$$L(b-a) \neq 0. \tag{4.110}$$

Note that (4.110) ensures that $L \neq 0$ and $a < b$. Combining this with the assumption that for all $x, y \in [a,b]^d$ it holds that

$$|f(x) - f(y)| \leq L\|x-y\|_1, \tag{4.111}$$

ensures that $L > 0$. Proposition 4.3.7 hence demonstrates that there exists $\mathbf{F} \in \mathbf{N}$ which satisfies that

(I) it holds that $\mathcal{I}(\mathbf{F}) = d$,

(II) it holds that $\mathcal{O}(\mathbf{F}) = 1$,

(III) it holds that $\mathcal{H}(\mathbf{F}) \leq \big(\lceil d\log_2\big(\frac{3dL(b-a)}{2\varepsilon}\big)\rceil + 1\big)\mathbb{1}_{(0,d/4)}\big(\frac{\varepsilon}{2L(b-a)}\big)$,

(IV) it holds that $\mathbb{D}_1(\mathbf{F}) \leq 2d\big(\frac{3dL(b-a)}{2\varepsilon}\big)^d\mathbb{1}_{(0,d/4)}\big(\frac{\varepsilon}{2L(b-a)}\big) + \mathbb{1}_{[d/4,\infty)}\big(\frac{\varepsilon}{2L(b-a)}\big)$,

(V) it holds for all $i \in \{2,3,4,\dots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3\big\lceil\big(\frac{3dL(b-a)}{2\varepsilon}\big)^d\frac{1}{2^{i-1}}\big\rceil$,

(VI) it holds that $\mathcal{P}(\mathbf{F}) \leq 35\big(\frac{3dL(b-a)}{2\varepsilon}\big)^{2d}d^2\mathbb{1}_{(0,d/4)}\big(\frac{\varepsilon}{2L(b-a)}\big) + (d+1)\mathbb{1}_{[d/4,\infty)}\big(\frac{\varepsilon}{2L(b-a)}\big)$,

(VII) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d}|f(x)|]\}$, and

(VIII) it holds that $\sup_{x\in[a,b]^d}|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 4.2.6). Observe that item (III) assures that

$$\begin{aligned}\mathcal{H}(\mathbf{F}) &\leq \big(d\big(\log_2\big(\tfrac{3dL(b-a)}{2}\big) + \log_2(\varepsilon^{-1})\big) + 2\big)\mathbb{1}_{(0,d/4)}\big(\tfrac{\varepsilon}{2L(b-a)}\big)\\ &\leq d\big(\max\big\{\log_2\big(\tfrac{3dL(b-a)}{2}\big), 0\big\} + \log_2(\varepsilon^{-1})\big) + 2.\end{aligned} \tag{4.112}$$

Furthermore, note that item (IV) ensures that

$$\begin{aligned}\mathbb{D}_1(\mathbf{F}) &\leq d\big(\tfrac{3d\max\{L(b-a),1\}}{\varepsilon}\big)^d\mathbb{1}_{(0,d/4)}\big(\tfrac{\varepsilon}{2L(b-a)}\big) + \mathbb{1}_{[d/4,\infty)}\big(\tfrac{\varepsilon}{2L(b-a)}\big)\\ &\leq \varepsilon^{-d}d(3d\max\{L(b-a),1\})^d.\end{aligned} \tag{4.113}$$

Moreover, observe that item (V) establishes that for all $i \in \{2,3,4,\dots\}$ it holds that

$$\mathbb{D}_i(\mathbf{F}) \leq 3\big(\big(\tfrac{3dL(b-a)}{2\varepsilon}\big)^d\tfrac{1}{2^{i-1}} + 1\big) \leq \varepsilon^{-d}3\big(\tfrac{(3dL(b-a))^d}{2^i} + 1\big). \tag{4.114}$$

In addition, note that item (VI) ensures that

$$\begin{aligned}\mathcal{P}(\mathbf{F}) &\leq 9\big(\tfrac{3d\max\{L(b-a),1\}}{\varepsilon}\big)^{2d} d^2 \mathbb{1}_{(0,d/4)}\big(\tfrac{\varepsilon}{2L(b-a)}\big) + (d+1)\mathbb{1}_{[d/4,\infty)}\big(\tfrac{\varepsilon}{2L(b-a)}\big) \\ &\leq \varepsilon^{-2d} 9\big(3d\max\{L(b-a),1\}\big)^{2d} d^2.\end{aligned} \tag{4.115}$$

Combining this, (4.112), (4.113), and (4.114) with items (I), (II), (VII), and (VIII) establishes items (i), (ii), (iii), (iv), (v), (vi), (vii), and (viii). The proof of Proposition 4.3.8 is thus complete. □

Corollary 4.3.9 (Implicit multi-dimensional ANN approximations with prescribed error tolerances and asymptotic parameter bounds)**.** *Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in [a,\infty)$ and let $f\colon [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that*

$$|f(x) - f(y)| \leq L\|x-y\|_1 \tag{4.116}$$

(cf. Definition 3.3.4). 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

$$\mathcal{H}(\mathbf{F}) \leq \mathfrak{C}(\log_2(\varepsilon^{-1}) + 1), \qquad \|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\big\{1, L, |a|, |b|, 2\big[\sup_{x\in[a,b]^d}|f(x)|\big]\big\}, \tag{4.117}$$

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \quad \sup_{x\in[a,b]^d}|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \tag{4.118}$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Corollary 4.3.9. Throughout this proof, let $\mathfrak{C} \in \mathbb{R}$ satisfy

$$\mathfrak{C} = 9\big(3d\max\{L(b-a),1\}\big)^{2d} d^2. \tag{4.119}$$

Observe that items (i), (ii), (iii), (vi), (vii), and (viii) in Proposition 4.3.8 and the fact that for all $\varepsilon \in (0,1]$ it holds that

$$\begin{aligned} d\big(\log_2\big(\max\big\{\tfrac{3dL(b-a)}{2},1\big\}\big) + \log_2(\varepsilon^{-1})\big) + 2 &\leq d\big(\max\big\{\tfrac{3dL(b-a)}{2},1\big\} + \log_2(\varepsilon^{-1})\big) + 2 \\ &\leq d\max\big\{3dL(b-a),1\big\} + 2 + d\log_2(\varepsilon^{-1}) \\ &\leq \mathfrak{C}(\log_2(\varepsilon^{-1}) + 1)\end{aligned} \tag{4.120}$$

imply that for every $\varepsilon \in (0,1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{H}(\mathbf{F}) \leq \mathfrak{C}(\log_2(\varepsilon^{-1}) + 1), \qquad \|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\big\{1, L, |a|, |b|, 2\big[\sup_{x\in[a,b]^d}|f(x)|\big]\big\}, \tag{4.121}$$

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \qquad \sup_{x\in[a,b]^d}|(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \tag{4.122}$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6). The proof of Corollary 4.3.9 is thus complete. □

Lemma 4.3.10 (Explicit estimates for vector norms). *Let $d \in \mathbb{N}$, $p, q \in (0, \infty]$ satisfy $p \leq q$. Then it holds for all $x \in \mathbb{R}^d$ that*

$$\|x\|_p \geq \|x\|_q \tag{4.123}$$

(cf. Definition 3.3.4).

Proof of Lemma 4.3.10. Throughout this proof, assume without loss of generality that $q < \infty$, let $e_1, e_2, \ldots, e_d \in \mathbb{R}^d$ satisfy $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$, $\ldots$, $e_d = (0, \ldots, 0, 1)$, let $r \in \mathbb{R}$ satisfy

$$r = p^{-1}q, \tag{4.124}$$

and let $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$ satisfy for all $i \in \{1, 2, \ldots, d\}$ that

$$y_i = |x_i|^p. \tag{4.125}$$

Note that (4.125), the fact that

$$y = \sum_{i=1}^{d} y_i e_i, \tag{4.126}$$

and the fact that for all $v, w \in \mathbb{R}^d$ it holds that

$$\|v + w\|_r \leq \|v\|_r + \|w\|_r \tag{4.127}$$

(cf. Definition 3.3.4) ensures that

$$\begin{aligned}
\|x\|_q &= \left[\sum_{i=1}^{d} |x_i|^q\right]^{1/q} = \left[\sum_{i=1}^{d} |x_i|^{pr}\right]^{1/q} = \left[\sum_{i=1}^{d} |y_i|^r\right]^{1/q} = \left[\sum_{i=1}^{d} |y_i|^r\right]^{1/(pr)} = \|y\|_r^{1/p} \\
&= \left\|\sum_{i=1}^{d} y_i e_i\right\|_r^{1/p} \leq \left[\sum_{i=1}^{d} \|y_i e_i\|_r\right]^{1/p} = \left[\sum_{i=1}^{d} |y_i| \|e_i\|_r\right]^{1/p} = \left[\sum_{i=1}^{d} |y_i|\right]^{1/p} \\
&= \|y\|_1^{1/p} = \|x\|_p.
\end{aligned} \tag{4.128}$$

This establishes (4.123). The proof of Lemma 4.3.10 is thus complete. □

Corollary 4.3.11 (Implicit multi-dimensional ANN approximations with prescribed error tolerances and asymptotic parameter bounds). *Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in [a, \infty)$ and let $f \colon [a, b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that*

$$|f(x) - f(y)| \leq L\|x - y\|_1 \tag{4.129}$$

(cf. Definition 3.3.4). Then there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $\varepsilon \in (0, 1]$ there exists

$\mathbf{F} \in \mathbf{N}$ *such that*

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \quad \sup_{x\in[a,b]^d} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \tag{4.130}$$

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

Proof of Corollary 4.3.11. Observe that Corollary 4.3.9 establishes (4.130). The proof of Corollary 4.3.11 is thus complete. □

4.4 Refined ANN approximations results for multi-dimensional functions

In Chapter 15 below we establish estimates for the *overall error* in the training of suitable rectified clipped ANNs (see Section 4.4.1 below) in the specific situation of GD-type optimization methods with many independent random initializations. Besides *optimization error* estimates from Part III and *generalization error* estimates from Part IV, for this overall error analysis we also employ suitable *approximation error* estimates with a somewhat more refined control on the architecture of the approximating ANNs than the approximation error estimates established in the previous sections of this chapter (cf., for instance, Corollaries 4.3.9 and 4.3.11 above). It is exactly the subject of this section to establish such refined approximation error estimates (see Proposition 4.4.12 below).

This section is specifically tailored to the requirements of the overall error analysis presented in Chapter 15 and does not offer much more significant insights into the approximation error analyses of ANNs than the content of the previous sections in this chapter. It can therefore be skipped at the first reading of this book and only needs to be considered when the reader is studying Chapter 15 in detail.

4.4.1 Rectified clipped ANNs

Definition 4.4.1 (Rectified clipped ANNs)**.** *Let* $L, \mathfrak{d} \in \mathbb{N}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $\mathbf{l} = (l_0, l_1, \ldots, l_L) \in \mathbb{N}^{L+1}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ *satisfy*

$$\mathfrak{d} \geq \sum_{k=1}^{L} l_k(l_{k-1} + 1). \tag{4.131}$$

Then we denote by $\mathcal{N}_{u,v}^{\theta,\mathbf{l}} \colon \mathbb{R}^{l_0} \to \mathbb{R}^{l_L}$ *the function which satisfies for all* $x \in \mathbb{R}^{l_0}$ *that*

$$\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(x) = \begin{cases} \left(\mathcal{N}_{\mathfrak{C}_{u,v,l_L}}^{\theta,l_0}\right)(x) & : L = 1 \\ \left(\mathcal{N}_{\mathfrak{R}_{l_1},\mathfrak{R}_{l_2},\ldots,\mathfrak{R}_{l_{L-1}},\mathfrak{C}_{u,v,l_L}}^{\theta,l_0}\right)(x) & : L > 1 \end{cases} \tag{4.132}$$

(cf. Definitions 1.1.3, 1.2.5, and 1.2.10).

Lemma 4.4.2. *Let* $\Phi \in \mathbf{N}$ *(cf. Definition 1.3.1). Then it holds for all* $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ *that*

$$\left(\mathcal{N}_{-\infty,\infty}^{\mathcal{I}(\Phi),\mathcal{D}(\Phi)}\right)(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x) \tag{4.133}$$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 4.4.1).

Proof of Lemma 4.4.2. Note that Proposition 1.3.10, (4.132), (1.27), and the fact that for all $d \in \mathbb{N}$ it holds that $\mathfrak{C}_{-\infty,\infty,d} = \mathrm{id}_{\mathbb{R}^d}$ prove (4.133) (cf. Definition 1.2.10). The proof of Lemma 4.4.2 is thus complete. □

4.4.2 Embedding ANNs in larger architectures

Lemma 4.4.3. *Let* $a \in C(\mathbb{R},\mathbb{R})$, $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_L \in \mathbb{N}$ *satisfy for all* $k \in \{1,2,\ldots,L\}$ *that* $\mathfrak{l}_0 = l_0$, $\mathfrak{l}_L = l_L$, *and* $\mathfrak{l}_k \geq l_k$, *for every* $k \in \{1,2,\ldots,L\}$ *let* $W_k = (W_{k,i,j})_{(i,j)\in\{1,2,\ldots,l_k\}\times\{1,2,\ldots,l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $\mathscr{W}_k = (\mathscr{W}_{k,i,j})_{(i,j)\in\{1,2,\ldots,\mathfrak{l}_k\}\times\{1,2,\ldots,\mathfrak{l}_{k-1}\}} \in \mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}}$, $B_k = (B_{k,i})_{i\in\{1,2,\ldots,l_k\}} \in \mathbb{R}^{l_k}$, $\mathscr{B}_k = (\mathscr{B}_{k,i})_{i\in\{1,2,\ldots,\mathfrak{l}_k\}} \in \mathbb{R}^{\mathfrak{l}_k}$, *assume for all* $k \in \{1,2,\ldots,L\}$, $i \in \{1,2,\ldots,l_k\}$, $j \in \mathbb{N} \cap (0, l_{k-1}]$ *that*

$$\mathscr{W}_{k,i,j} = W_{k,i,j} \qquad \text{and} \qquad \mathscr{B}_{k,i} = B_{k,i}, \tag{4.134}$$

and assume for all $k \in \{1,2,\ldots,L\}$, $i \in \{1,2,\ldots,l_k\}$, $j \in \mathbb{N} \cap (l_{k-1}, \mathfrak{l}_{k-1}+1)$ *that* $\mathscr{W}_{k,i,j} = 0$. *Then*

$$\mathcal{R}_a^{\mathbf{N}}\big(((W_1,B_1),(W_2,B_2),\ldots,(W_L,B_L))\big) = \mathcal{R}_a^{\mathbf{N}}\big(((\mathscr{W}_1,\mathscr{B}_1),(\mathscr{W}_2,\mathscr{B}_2),\ldots,(\mathscr{W}_L,\mathscr{B}_L))\big) \tag{4.135}$$

(cf. Definition 1.3.4).

Proof of Lemma 4.4.3. Throughout this proof, let $\pi_k \colon \mathbb{R}^{\mathfrak{l}_k} \to \mathbb{R}^{l_k}$, $k \in \{0,1,\ldots,L\}$, satisfy for all $k \in \{0,1,\ldots,L\}$, $x = (x_1, x_2, \ldots, x_{\mathfrak{l}_k})$ that

$$\pi_k(x) = (x_1, x_2, \ldots, x_{l_k}). \tag{4.136}$$

Observe that the assumption that $\mathfrak{l}_0 = l_0$ and $\mathfrak{l}_L = l_L$ implies that

$$\mathcal{R}_a^{\mathbf{N}}\big(((W_1,B_1),(W_2,B_2),\ldots,(W_L,B_L))\big) \in C(\mathbb{R}^{\mathfrak{l}_0},\mathbb{R}^{\mathfrak{l}_L}) \tag{4.137}$$

(cf. Definition 1.3.4). Furthermore, note that the assumption that for all $k \in \{1, 2, \ldots, l\}$, $i \in \{1, 2, \ldots, l_k\}$, $j \in \mathbb{N} \cap (l_{k-1}, \mathfrak{l}_{k-1} + 1)$ it holds that $\mathscr{W}_{k,i,j} = 0$ shows that for all $k \in \{1, 2, \ldots, L\}$, $x = (x_1, \ldots, x_{\mathfrak{l}_{k-1}}) \in \mathbb{R}^{\mathfrak{l}_{k-1}}$ it holds that

$$\begin{aligned}
&\pi_k(\mathscr{W}_k x + \mathscr{B}_k) \\
&= \left(\left[\sum_{i=1}^{l_{k-1}} \mathscr{W}_{k,1,i} x_i \right] + \mathscr{B}_{k,1}, \left[\sum_{i=1}^{l_{k-1}} \mathscr{W}_{k,2,i} x_i \right] + \mathscr{B}_{k,2}, \ldots, \left[\sum_{i=1}^{l_{k-1}} \mathscr{W}_{k,l_k,i} x_i \right] + \mathscr{B}_{k,l_k} \right) \\
&= \left(\left[\sum_{i=1}^{l_{k-1}} \mathscr{W}_{k,1,i} x_i \right] + \mathscr{B}_{k,1}, \left[\sum_{i=1}^{l_{k-1}} \mathscr{W}_{k,2,i} x_i \right] + \mathscr{B}_{k,2}, \ldots, \left[\sum_{i=1}^{l_{k-1}} \mathscr{W}_{k,l_k,i} x_i \right] + \mathscr{B}_{k,l_k} \right).
\end{aligned} \tag{4.138}$$

Combining this with the assumption that for all $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, l_k\}$, $j \in \mathbb{N} \cap (0, l_{k-1}]$ it holds that $\mathscr{W}_{k,i,j} = W_{k,i,j}$ and $\mathscr{B}_{k,i} = B_{k,i}$ demonstrates that for all $k \in \{1, 2, \ldots, L\}$, $x = (x_1, \ldots, x_{\mathfrak{l}_{k-1}}) \in \mathbb{R}^{\mathfrak{l}_{k-1}}$ it holds that

$$\begin{aligned}
&\pi_k(\mathscr{W}_k x + \mathscr{B}_k) \\
&= \left(\left[\sum_{i=1}^{l_{k-1}} W_{k,1,i} x_i \right] + B_{k,1}, \left[\sum_{i=1}^{l_{k-1}} W_{k,2,i} x_i \right] + B_{k,2}, \ldots, \left[\sum_{i=1}^{l_{k-1}} W_{k,l_k,i} x_i \right] + B_{k,l_k} \right) \\
&= W_k \pi_{k-1}(x) + B_k.
\end{aligned} \tag{4.139}$$

Therefore, we obtain that for all $x_0 \in \mathbb{R}^{\mathfrak{l}_0}$, $x_1 \in \mathbb{R}^{\mathfrak{l}_1}, \ldots, x_{L-1} \in \mathbb{R}^{\mathfrak{l}_{L-1}}$, $k \in \mathbb{N} \cap (0, L)$ with $\forall\, m \in \mathbb{N} \cap (0, L)\colon x_m = \mathfrak{M}_{a,\mathfrak{l}_m}(\mathscr{W}_m x_{m-1} + \mathscr{B}_m)$ it holds that

$$\pi_k(x_k) = \mathfrak{M}_{a,l_k}(\pi_k(\mathscr{W}_k x_{k-1} + \mathscr{B}_k)) = \mathfrak{M}_{a,l_k}(W_k \pi_{k-1}(x_{k-1}) + B_k) \tag{4.140}$$

(cf. Definition 1.2.1). Induction, the assumption that $l_0 = \mathfrak{l}_0$ and $l_L = \mathfrak{l}_L$, and (4.139) hence ensure that for all $x_0 \in \mathbb{R}^{\mathfrak{l}_0}$, $x_1 \in \mathbb{R}^{\mathfrak{l}_1}, \ldots, x_{L-1} \in \mathbb{R}^{\mathfrak{l}_{L-1}}$ with $\forall\, k \in \mathbb{N} \cap (0, L)\colon x_k = \mathfrak{M}_{a,\mathfrak{l}_k}(\mathscr{W}_k x_{k-1} + \mathscr{B}_k)$ it holds that

$$\begin{aligned}
&\big(\mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)))\big)(x_0) \\
&= \big(\mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)))\big)(\pi_0(x_0)) \\
&= W_L \pi_{L-1}(x_{L-1}) + B_L \\
&= \pi_L(\mathscr{W}_L x_{L-1} + \mathscr{B}_L) = \mathscr{W}_L x_{L-1} + \mathscr{B}_L \\
&= \big(\mathcal{R}_a^{\mathbf{N}}(((\mathscr{W}_1, \mathscr{B}_1), (\mathscr{W}_2, \mathscr{B}_2), \ldots, (\mathscr{W}_L, \mathscr{B}_L)))\big)(x_0).
\end{aligned} \tag{4.141}$$

The proof of Lemma 4.4.3 is thus complete. □

Lemma 4.4.4. *Let $a \in C(\mathbb{R}, \mathbb{R})$, $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_L \in \mathbb{N}$ satisfy for all $k \in \{1, 2, \ldots, L\}$ that*

$$\mathfrak{l}_0 = l_0, \qquad \mathfrak{l}_L = l_L, \qquad \text{and} \qquad \mathfrak{l}_k \geq l_k \tag{4.142}$$

and let $\Phi \in \mathbf{N}$ satisfy $\mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_L)$ (cf. Definition 1.3.1). Then there exists $\Psi \in \mathbf{N}$ such that

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_L), \qquad \|\mathcal{T}(\Psi)\|_\infty = \|\mathcal{T}(\Phi)\|_\infty, \qquad \text{and} \qquad \mathcal{R}^{\mathbf{N}}_a(\Psi) = \mathcal{R}^{\mathbf{N}}_a(\Phi) \tag{4.143}$$

(cf. Definitions 1.3.4, 1.3.6, and 3.3.4).

Proof of Lemma 4.4.4. Throughout this proof, let $B_k = (B_{k,i})_{i \in \{1,2,\ldots,l_k\}} \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \ldots, L\}$, and $W_k = (W_{k,i,j})_{(i,j) \in \{1,2,\ldots,l_k\} \times \{1,2,\ldots,l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $k \in \{1, 2, \ldots, L\}$, satisfy

$$\Phi = ((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)) \tag{4.144}$$

and let $\mathfrak{W}_k = (\mathfrak{W}_{k,i,j})_{(i,j) \in \{1,2,\ldots,\mathfrak{l}_k\} \times \{1,2,\ldots,\mathfrak{l}_{k-1}\}} \in \mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}}$, $k \in \{1, 2, \ldots, L\}$, and $\mathfrak{B}_k = (\mathfrak{B}_{k,i})_{i \in \{1,2,\ldots,\mathfrak{l}_k\}} \in \mathbb{R}^{\mathfrak{l}_k}$, $k \in \{1, 2, \ldots, L\}$, satisfy for all $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, \mathfrak{l}_k\}$, $j \in \{1, 2, \ldots, \mathfrak{l}_{k-1}\}$ that

$$\mathfrak{W}_{k,i,j} = \begin{cases} W_{k,i,j} & : (i \leq l_k) \wedge (j \leq l_{k-1}) \\ 0 & : (i > l_k) \vee (j > l_{k-1}) \end{cases} \qquad \text{and} \qquad \mathfrak{B}_{k,i} = \begin{cases} B_{k,i} & : i \leq l_k \\ 0 & : i > l_k. \end{cases} \tag{4.145}$$

Observe that (1.78) establishes that $((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L)) \in \big(\bigtimes_{i=1}^{L} (\mathbb{R}^{\mathfrak{l}_i \times \mathfrak{l}_{i-1}} \times \mathbb{R}^{\mathfrak{l}_i})\big) \subseteq \mathbf{N}$ and

$$\mathcal{D}\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big) = (\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_L). \tag{4.146}$$

Furthermore, note that Lemma 1.3.9 and (4.145) prove that

$$\|\mathcal{T}\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big)\|_\infty = \|\mathcal{T}(\Phi)\|_\infty \tag{4.147}$$

(cf. Definitions 1.3.6 and 3.3.4). Moreover, observe that Lemma 4.4.3 implies that

$$\begin{aligned} \mathcal{R}^{\mathbf{N}}_a(\Phi) &= \mathcal{R}^{\mathbf{N}}_a\big(((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L))\big) \\ &= \mathcal{R}^{\mathbf{N}}_a\big(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_L, \mathfrak{B}_L))\big) \end{aligned} \tag{4.148}$$

(cf. Definition 1.3.4). The proof of Lemma 4.4.4 is thus complete. □

Lemma 4.4.5. *Let $L, \mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}} \in \mathbb{N}$, $\Phi_1 = ((W_1, B_1), (W_2, B_2), \ldots, (W_L, B_L)) \in \big(\bigtimes_{k=1}^{L} (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})\big)$, $\Phi_2 = ((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \ldots, (\mathfrak{W}_{\mathfrak{L}}, \mathfrak{B}_{\mathfrak{L}})) \in \big(\bigtimes_{k=1}^{\mathfrak{L}} (\mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}} \times \mathbb{R}^{\mathfrak{l}_k})\big)$. Then*

$$\|\mathcal{T}(\Phi_1 \bullet \Phi_2)\|_\infty \leq \max\big\{\|\mathcal{T}(\Phi_1)\|_\infty, \|\mathcal{T}(\Phi_2)\|_\infty, \big\|\mathcal{T}\big(((W_1 \mathfrak{W}_{\mathfrak{L}}, W_1 \mathfrak{B}_{\mathfrak{L}} + B_1))\big)\big\|_\infty\big\} \tag{4.149}$$

(cf. Definitions 1.3.6, 2.1.1, and 3.3.4).

Proof of Lemma 4.4.5. Note that (2.2) and Lemma 1.3.9 establish (4.149). The proof of Lemma 4.4.5 is thus complete. □

Lemma 4.4.6. *Let $d, L \in \mathbb{N}$, $\Phi \in \mathbf{N}$ satisfy $L \geq \mathcal{L}(\Phi)$ and $d = \mathcal{O}(\Phi)$ (cf. Definition 1.3.1). Then*

$$\|\mathcal{T}(\mathcal{E}_{L,\mathfrak{I}_d}(\Phi))\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \tag{4.150}$$

(cf. Definitions 1.3.6, 2.2.6, 2.2.9, and 3.3.4).

Proof of Lemma 4.4.6. Throughout this proof, assume without loss of generality that $L > \mathcal{L}(\Phi)$ and let $l_0, l_1, \ldots, l_{L-\mathcal{L}(\Phi)+1} \in \mathbb{N}$ satisfy

$$(l_0, l_1, \ldots, l_{L-\mathcal{L}(\Phi)+1}) = (d, 2d, 2d, \ldots, 2d, d). \tag{4.151}$$

Observe that Lemma 2.2.7 shows that $\mathcal{D}(\mathfrak{I}_d) = (d, 2d, d) \in \mathbb{N}^3$ (cf. Definition 2.2.6). Item (i) in Lemma 2.2.10 therefore demonstrates that

$$\begin{aligned} &\mathcal{L}((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))}) = L - \mathcal{L}(\Phi) + 1 \\ \text{and} \qquad &\mathcal{D}((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))}) = (l_0, l_1, \ldots, l_{L-\mathcal{L}(\Phi)+1}) \in \mathbb{N}^{L-\mathcal{L}(\Phi)+2} \end{aligned} \tag{4.152}$$

(cf. Definition 2.1.1). This ensures that there exist $W_k \in \mathbb{R}^{l_k \times l_{k-1}}$, $k \in \{1, 2, \ldots, L-\mathcal{L}(\Phi)+1\}$, and $B_k \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \ldots, L - \mathcal{L}(\Phi) + 1\}$, which satisfy

$$(\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))} = ((W_1, B_1), (W_2, B_2), \ldots, (W_{L-\mathcal{L}(\Phi)+1}, B_{L-\mathcal{L}(\Phi)+1})). \tag{4.153}$$

Furthermore, note that (2.44), (2.70), (2.71), (2.2), and (2.41) prove that

$$W_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{(2d)\times d} \tag{4.154}$$

$$\text{and} \qquad W_{L-\mathcal{L}(\Phi)+1} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{R}^{d\times(2d)}.$$

Moreover, observe that (2.44), (2.70), (2.71), (2.2), and (2.41) imply that for all $k \in \mathbb{N} \cap (1, L - \mathcal{L}(\Phi) + 1)$ it holds that

$$\begin{aligned} W_k &= \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix}}_{\in\mathbb{R}^{(2d)\times d}} \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{\in\mathbb{R}^{d\times(2d)}} \\ &= \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(2d)\times(2d)}. \end{aligned} \tag{4.155}$$

In addition, note that (2.70), (2.71), (2.44), (2.41), and (2.2) establish that for all $k \in \mathbb{N} \cap [1, L - \mathcal{L}(\Phi)]$ it holds that

$$B_k = 0 \in \mathbb{R}^{2d} \qquad \text{and} \qquad B_{L-\mathcal{L}(\Phi)+1} = 0 \in \mathbb{R}^d. \tag{4.156}$$

Combining this, (4.154), and (4.155) shows that

$$\left\| \mathcal{T}\left((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))} \right) \right\|_\infty = 1 \tag{4.157}$$

(cf. Definitions 1.3.6 and 3.3.4). Next observe that (4.154) demonstrates that for all $k \in \mathbb{N}$, $\mathfrak{W} = (w_{i,j})_{(i,j)\in\{1,2,\ldots,d\}\times\{1,2,\ldots,k\}} \in \mathbb{R}^{d\times k}$ it holds that

$$W_1\mathfrak{W} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,k} \\ -w_{1,1} & -w_{1,2} & \cdots & -w_{1,k} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,k} \\ -w_{2,1} & -w_{2,2} & \cdots & -w_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{d,1} & w_{d,2} & \cdots & w_{d,k} \\ -w_{d,1} & -w_{d,2} & \cdots & -w_{d,k} \end{pmatrix} \in \mathbb{R}^{(2d)\times k}. \tag{4.158}$$

Furthermore, note that (4.154) and (4.156) ensure that for all $\mathfrak{B} = (b_1, b_2, \ldots, b_d) \in \mathbb{R}^d$ it holds that

$$W_1\mathfrak{B} + B_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} = \begin{pmatrix} b_1 \\ -b_1 \\ b_2 \\ -b_2 \\ \vdots \\ b_d \\ -b_d \end{pmatrix} \in \mathbb{R}^{2d}. \tag{4.159}$$

Combining this with (4.158) proves that for all $k \in \mathbb{N}$, $\mathfrak{W} \in \mathbb{R}^{d\times k}$, $\mathfrak{B} \in \mathbb{R}^d$ it holds that

$$\big\|\mathcal{T}\big(((W_1\mathfrak{W}, W_1\mathfrak{B} + B_1))\big)\big\|_\infty = \big\|\mathcal{T}\big(((\mathfrak{W}, \mathfrak{B}))\big)\big\|_\infty. \tag{4.160}$$

This, Lemma 4.4.5, and (4.157) imply that

$$\begin{aligned} &\|\mathcal{T}(\mathcal{E}_{L,\mathfrak{I}_d}(\Phi))\|_\infty = \big\|\mathcal{T}\big(((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))}) \bullet \Phi\big)\big\|_\infty \\ &\leq \max\big\{\big\|\mathcal{T}\big((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))}\big)\big\|_\infty, \|\mathcal{T}(\Phi)\|_\infty\big\} = \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \end{aligned} \tag{4.161}$$

(cf. Definition 2.2.9). The proof of Lemma 4.4.6 is thus complete. □

Lemma 4.4.7. *Let $L, \mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_\mathfrak{L} \in \mathbb{N}$ satisfy*

$$\mathfrak{L} \geq L, \qquad \mathfrak{l}_0 = l_0, \qquad \text{and} \qquad \mathfrak{l}_\mathfrak{L} = l_L, \tag{4.162}$$

assume for all $i \in \mathbb{N} \cap [0, L)$ that $\mathfrak{l}_i \geq l_i$, assume for all $i \in \mathbb{N} \cap (L-1, \mathfrak{L})$ that $\mathfrak{l}_i \geq 2l_L$, and let $\Phi \in \mathbf{N}$ satisfy $\mathcal{D}(\Phi) = (l_0, l_1, \ldots, l_L)$ (cf. Definition 1.3.1). Then there exists

$\Psi \in \mathbf{N}$ *such that*

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}}), \quad \|\mathcal{T}(\Psi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\}, \quad \textit{and} \quad \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Psi) = \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi) \tag{4.163}$$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 3.3.4).

Proof of Lemma 4.4.7. Throughout this proof, let $\Xi \in \mathbf{N}$ satisfy $\Xi = \mathcal{E}_{\mathfrak{L}, \mathfrak{I}_{l_L}}(\Phi)$ (cf. Definitions 2.2.6 and 2.2.9). Observe that item (i) in Lemma 2.2.7 establishes that $\mathcal{D}(\mathfrak{I}_{l_L}) = (l_L, 2l_L, l_L) \in \mathbb{N}^3$. Combining this with Lemma 2.2.12 shows that $\mathcal{D}(\Xi) \in \mathbb{N}^{\mathfrak{L}+1}$ and

$$\mathcal{D}(\Xi) = \begin{cases} (l_0, l_1, \ldots, l_L) & : \mathfrak{L} = L \\ (l_0, l_1, \ldots, l_{L-1}, 2l_L, 2l_L, \ldots, 2l_L, l_L) & : \mathfrak{L} > L. \end{cases} \tag{4.164}$$

Furthermore, note that Lemma 4.4.6 (applied with $d \curvearrowleft l_L$, $L \curvearrowleft \mathfrak{L}$, $\Phi \curvearrowleft \Phi$ in the notation of Lemma 4.4.6) demonstrates that

$$\|\mathcal{T}(\Xi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \tag{4.165}$$

(cf. Definitions 1.3.6 and 3.3.4). Moreover, observe that item (ii) in Lemma 2.2.7 ensures that for all $x \in \mathbb{R}^{l_L}$ it holds that

$$(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathfrak{I}_{l_L}))(x) = x \tag{4.166}$$

(cf. Definitions 1.2.4 and 1.3.4). This and item (ii) in Lemma 2.2.11 prove that

$$\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Xi) = \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi). \tag{4.167}$$

In addition, note that (4.164), the assumption that for all $i \in [0, L)$ it holds that $\mathfrak{l}_0 = l_0$, $\mathfrak{l}_{\mathfrak{L}} = l_L$, and $\mathfrak{l}_i \leq l_i$, the assumption that for all $i \in \mathbb{N} \cap (L-1, \mathfrak{L})$ it holds that $\mathfrak{l}_i \geq 2l_L$, and Lemma 4.4.4 (applied with $a \curvearrowleft \mathfrak{r}$, $L \curvearrowleft \mathfrak{L}$, $(l_0, l_1, \ldots, l_L) \curvearrowleft \mathcal{D}(\Xi)$, $(\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}}) \curvearrowleft (\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}})$, $\Phi \curvearrowleft \Xi$ in the notation of Lemma 4.4.4) prove that there exists $\Psi \in \mathbf{N}$ such that

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}}), \qquad \|\mathcal{T}(\Psi)\|_\infty = \|\mathcal{T}(\Xi)\|_\infty, \qquad \text{and} \qquad \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Psi) = \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Xi). \tag{4.168}$$

Combining this with (4.165) and (4.167) establishes (4.163). The proof of Lemma 4.4.7 is thus complete. □

Lemma 4.4.8. *Let $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $L, \mathfrak{L}, d, \mathfrak{d} \in \mathbb{N}$, $\theta \in \mathbb{R}^d$, $l_0, l_1, \ldots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}} \in \mathbb{N}$ satisfy that*

$$d \geq \textstyle\sum_{i=1}^{L} l_i(l_{i-1}+1), \quad \mathfrak{d} \geq \textstyle\sum_{i=1}^{\mathfrak{L}} \mathfrak{l}_i(\mathfrak{l}_{i-1}+1), \quad \mathfrak{L} \geq L, \quad \mathfrak{l}_0 = l_0, \quad \textit{and} \quad \mathfrak{l}_{\mathfrak{L}} = l_L, \tag{4.169}$$

assume for all $i \in \mathbb{N} \cap [0, L)$ that $\mathfrak{l}_i \geq l_i$, and assume for all $i \in \mathbb{N} \cap (L-1, \mathfrak{L})$ that

$\mathfrak{l}_i \geq 2l_L$. *Then there exists* $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ *such that*

$$\|\vartheta\|_\infty \leq \max\{1, \|\theta\|_\infty\} \qquad \text{and} \qquad \mathscr{N}_{u,v}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_{\mathfrak{L}})} = \mathscr{N}_{u,v}^{\theta,(l_0,l_1,\ldots,l_L)} \tag{4.170}$$

(cf. Definitions 3.3.4 and 4.4.1).

Proof of Lemma 4.4.8. Throughout this proof, let $\eta_1, \eta_2, \ldots, \eta_d \in \mathbb{R}$ satisfy

$$\theta = (\eta_1, \eta_2, \ldots, \eta_d) \tag{4.171}$$

and let $\Phi \in \left(\bigtimes_{i=1}^{L} \mathbb{R}^{l_i \times l_{i-1}} \times \mathbb{R}^{l_i}\right)$ satisfy

$$\mathcal{T}(\Phi) = (\eta_1, \eta_2, \ldots, \eta_{\mathcal{P}(\Phi)}) \tag{4.172}$$

(cf. Definitions 1.3.1 and 1.3.6). Observe that Lemma 4.4.7 implies that there exists $\Psi \in \mathbf{N}$ which satisfies

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}}), \quad \|\mathcal{T}(\Psi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\}, \quad \text{and} \quad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi) = \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi) \tag{4.173}$$

(cf. Definitions 1.2.4, 1.3.4, and 3.3.4). Next let $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$(\vartheta_1, \vartheta_2, \ldots, \vartheta_{\mathcal{P}(\Psi)}) = \mathcal{T}(\Psi) \qquad \text{and} \qquad \forall\, i \in \mathbb{N} \cap (\mathcal{P}(\Psi), \mathfrak{d}+1)\colon \vartheta_i = 0. \tag{4.174}$$

Note that (4.171), (4.172), (4.173), and (4.174) show that

$$\|\vartheta\|_\infty = \|\mathcal{T}(\Psi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \leq \max\{1, \|\theta\|_\infty\}. \tag{4.175}$$

Furthermore, observe that Lemma 4.4.2 and (4.172) demonstrate that for all $x \in \mathbb{R}^{l_0}$ it holds that

$$\left(\mathscr{N}_{-\infty,\infty}^{\theta,(l_0,l_1,\ldots,l_L)}\right)(x) = \left(\mathscr{N}_{-\infty,\infty}^{\mathcal{T}(\Phi),\mathcal{D}(\Phi)}\right)(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x) \tag{4.176}$$

(cf. Definition 4.4.1). Moreover, note that Lemma 4.4.2, (4.173), and (4.174) ensure that for all $x \in \mathbb{R}^{\mathfrak{l}_0}$ it holds that

$$\left(\mathscr{N}_{-\infty,\infty}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_{\mathfrak{L}})}\right)(x) = \left(\mathscr{N}_{-\infty,\infty}^{\mathcal{T}(\Psi),\mathcal{D}(\Psi)}\right)(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi))(x). \tag{4.177}$$

Combining this and (4.176) with (4.173) and the assumption that $\mathfrak{l}_0 = l_0$ and $\mathfrak{l}_{\mathfrak{L}} = l_L$ proves that

$$\mathscr{N}_{-\infty,\infty}^{\theta,(l_0,l_1,\ldots,l_L)} = \mathscr{N}_{-\infty,\infty}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_{\mathfrak{L}})}. \tag{4.178}$$

Hence, we obtain that

$$\mathscr{N}_{u,v}^{\theta,(l_0,l_1,\ldots,l_L)} = \mathfrak{C}_{u,v,l_L} \circ \mathscr{N}_{-\infty,\infty}^{\theta,(l_0,l_1,\ldots,l_L)} = \mathfrak{C}_{u,v,\mathfrak{l}_{\mathfrak{L}}} \circ \mathscr{N}_{-\infty,\infty}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_{\mathfrak{L}})} = \mathscr{N}_{u,v}^{\vartheta,(\mathfrak{l}_0,\mathfrak{l}_1,\ldots,\mathfrak{l}_{\mathfrak{L}})} \tag{4.179}$$

(cf. Definition 1.2.10). This and (4.175) establish (4.170). The proof of Lemma 4.4.8 is thus complete. □

4.4.3 Approximation through ANNs with variable architectures

Corollary 4.4.9. *Let $d, K, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \ldots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, $L \in [0, \infty)$ satisfy that*

$$\mathbf{L} \geq \lceil \log_2(K) \rceil + 2, \quad \mathbf{l}_0 = d, \quad \mathbf{l}_{\mathbf{L}} = 1, \quad \mathbf{l}_1 \geq 2dK, \quad \text{and} \quad \mathbf{d} \geq \textstyle\sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1), \tag{4.180}$$

assume for all $i \in \mathbb{N} \cap (1, \mathbf{L})$ that $\mathbf{l}_i \geq 3\lceil \frac{K}{2^{i-1}} \rceil$, let $E \subseteq \mathbb{R}^d$ be a set, let $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in E$, and let $f \colon E \to \mathbb{R}$ satisfy for all $x, y \in E$ that $|f(x) - f(y)| \leq L\|x - y\|_1$ (cf. Definitions 3.3.4 and 4.2.6). Then there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k \in \{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1,2,\ldots,K\}} |f(\mathfrak{x}_k)|\} \tag{4.181}$$

and

$$\sup_{x \in E} \big| f(x) - \mathscr{N}^{\theta,\mathbf{l}}_{-\infty,\infty}(x) \big| \leq 2L \big[\sup_{x \in E} \big(\inf_{k \in \{1,2,\ldots,K\}} \|x - \mathfrak{x}_k\|_1 \big) \big] \tag{4.182}$$

(cf. Definition 4.4.1).

Proof of Corollary 4.4.9. Throughout this proof, let $\mathfrak{y} \in \mathbb{R}^K$, $\Phi \in \mathbf{N}$ satisfy $\mathfrak{y} = (f(\mathfrak{x}_1), f(\mathfrak{x}_2), \ldots, f(\mathfrak{x}_K))$ and

$$\Phi = \mathbb{M}_K \bullet \mathbf{A}_{-L\mathrm{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K \big(\mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_2}, \ldots, \mathbb{L}_d \bullet \mathbf{A}_{\mathrm{I}_d, -\mathfrak{x}_K} \big) \bullet \mathbb{T}_{d,K} \tag{4.183}$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, 2.2.1, 2.3.1, 2.4.6, 4.2.1, and 4.2.5). Observe that Lemma 4.2.9 and Proposition 4.3.1 imply that

(I) it holds that $\mathcal{L}(\Phi) = \lceil \log_2(K) \rceil + 2$,

(II) it holds that $\mathcal{I}(\Phi) = d$,

(III) it holds that $\mathcal{O}(\Phi) = 1$,

(IV) it holds that $\mathbb{D}_1(\Phi) = 2dK$,

(V) it holds for all $i \in \{2, 3, \ldots, \mathcal{L}(\Phi) - 1\}$ that $\mathbb{D}_i(\Phi) \leq 3\lceil \frac{K}{2^{i-1}} \rceil$,

(VI) it holds that $\|\mathcal{T}(\Phi)\|_\infty \leq \max\{1, L, \max_{k \in \{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1,2,\ldots,K\}} |f(\mathfrak{x}_k)|\}$, and

(VII) it holds that $\sup_{x \in E} |f(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x)| \leq 2L \big[\sup_{x \in E} \big(\inf_{k \in \{1,2,\ldots,K\}} \|x - \mathfrak{x}_k\|_1 \big) \big]$

(cf. Definitions 1.2.4, 1.3.4, and 1.3.6). Furthermore, note that the fact that $\mathbf{L} \geq \lceil \log_2(K) \rceil + 2 = \mathcal{L}(\Phi)$, the fact that $\mathbf{l}_0 = d = \mathbb{D}_0(\Phi)$, the fact that $\mathbf{l}_1 \geq 2dK = \mathbb{D}_1(\Phi)$, the fact that for all $i \in \{1, 2, \ldots, \mathcal{L}(\Phi) - 1\} \backslash \{1\}$ it holds that $\mathbf{l}_i \geq 3\lceil \frac{K}{2^{i-1}} \rceil \geq \mathbb{D}_i(\Phi)$, the fact that for all $i \in \mathbb{N} \cap (\mathcal{L}(\Phi) - 1, \mathbf{L})$ it holds that $\mathbf{l}_i \geq 3\lceil \frac{K}{2^{i-1}} \rceil \geq 2 = 2\mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)$, the fact that $\mathbf{l}_{\mathbf{L}} = 1 = \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)$, and Lemma 4.4.8 show that there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ which satisfies that

$$\|\theta\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \qquad \text{and} \qquad \mathscr{N}^{\theta,(\mathbf{l}_0,\mathbf{l}_1,\ldots,\mathbf{l}_{\mathbf{L}})}_{-\infty,\infty} = \mathscr{N}^{\mathcal{T}(\Phi),\mathcal{D}(\Phi)}_{-\infty,\infty}. \tag{4.184}$$

This and item (VI) demonstrate that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k\in\{1,2,\ldots,K\}}\|\mathfrak{x}_k\|_\infty, 2\max_{k\in\{1,2,\ldots,K\}}|f(\mathfrak{x}_k)|\}. \tag{4.185}$$

Moreover, observe that (4.184), Lemma 4.4.2, and item (VII) ensure that

$$\begin{aligned}\sup_{x\in E}\big|f(x) - \mathscr{N}^{\theta,(\mathbf{l}_0,\mathbf{l}_1,\ldots,\mathbf{l}_\mathbf{L})}_{-\infty,\infty}(x)\big| &= \sup_{x\in E}\big|f(x) - \mathscr{N}^{\mathcal{T}(\Phi),\mathcal{D}(\Phi)}_{-\infty,\infty}(x)\big| \\ &= \sup_{x\in E}\big|f(x) - (\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\Phi))(x)\big| \\ &\leq 2L\big[\sup_{x\in E}\big(\inf_{k\in\{1,2,\ldots,K\}}\|x-\mathfrak{x}_k\|_1\big)\big]\end{aligned} \tag{4.186}$$

(cf. Definition 4.4.1). The proof of Corollary 4.4.9 is thus complete. □

Corollary 4.4.10. *Let $d, K, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \ldots, \mathbf{l}_\mathbf{L}) \in \mathbb{N}^{\mathbf{L}+1}$, $L \in [0,\infty)$, $u \in [-\infty,\infty)$, $v \in (u,\infty]$ satisfy that*

$$\mathbf{L} \geq \lceil \log_2 K\rceil + 2, \quad \mathbf{l}_0 = d, \quad \mathbf{l}_\mathbf{L} = 1, \quad \mathbf{l}_1 \geq 2dK, \quad \text{and} \quad \mathbf{d} \geq \textstyle\sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1}+1), \tag{4.187}$$

assume for all $i \in \mathbb{N}\cap(1,\mathbf{L})$ that $\mathbf{l}_i \geq 3\lceil \frac{K}{2^{i-1}}\rceil$, let $E \subseteq \mathbb{R}^d$ be a set, let $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in E$, and let $f\colon E \to ([u,v]\cap\mathbb{R})$ satisfy for all $x, y \in E$ that $|f(x) - f(y)| \leq L\|x-y\|_1$ (cf. Definitions 3.3.4 and 4.2.6). Then there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k\in\{1,2,\ldots,K\}}\|\mathfrak{x}_k\|_\infty, 2\max_{k\in\{1,2,\ldots,K\}}|f(\mathfrak{x}_k)|\} \tag{4.188}$$

and

$$\sup_{x\in E}\big|f(x) - \mathscr{N}^{\theta,\mathbf{l}}_{u,v}(x)\big| \leq 2L\big[\sup_{x\in E}\big(\inf_{k\in\{1,2,\ldots,K\}}\|x-\mathfrak{x}_k\|_1\big)\big]. \tag{4.189}$$

(cf. Definition 4.4.1).

Proof of Corollary 4.4.10. Note that Corollary 4.4.9 proves that there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k\in\{1,2,\ldots,K\}}\|\mathfrak{x}_k\|_\infty, 2\max_{k\in\{1,2,\ldots,K\}}|f(\mathfrak{x}_k)|\} \tag{4.190}$$

and

$$\sup_{x\in E}\big|f(x) - \mathscr{N}^{\theta,\mathbf{l}}_{-\infty,\infty}(x)\big| \leq 2L\big[\sup_{x\in E}\big(\inf_{k\in\{1,2,\ldots,K\}}\|x-\mathfrak{x}_k\|_1\big)\big]. \tag{4.191}$$

Furthermore, observe that the assumption that $f(E) \subseteq [u,v]$ establishes that for all $x \in E$ it holds that

$$f(x) = \mathfrak{c}_{u,v}(f(x)) \tag{4.192}$$

(cf. Definitions 1.2.9 and 4.4.1). The fact that for all $x, y \in \mathbb{R}$ it holds that $|\mathfrak{c}_{u,v}(x) - \mathfrak{c}_{u,v}(y)| \leq |x-y|$ and (4.191) therefore imply that

$$\begin{aligned}&\sup_{x\in E}\big|f(x) - \mathscr{N}^{\theta,\mathbf{l}}_{u,v}(x)\big| = \sup_{x\in E}|\mathfrak{c}_{u,v}(f(x)) - \mathfrak{c}_{u,v}(\mathscr{N}^{\theta,\mathbf{l}}_{-\infty,\infty}(x))| \\ &\leq \sup_{x\in E}\big|f(x) - \mathscr{N}^{\theta,\mathbf{l}}_{-\infty,\infty}(x)\big| \leq 2L\big[\sup_{x\in E}\big(\inf_{k\in\{1,2,\ldots,K\}}\|x-\mathfrak{x}_k\|_1\big)\big].\end{aligned} \tag{4.193}$$

The proof of Corollary 4.4.10 is thus complete. □

4.4.4 Refined convergence rates for the approximation error

Lemma 4.4.11. *Let $d, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \ldots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, assume $\mathbf{l}_0 = d$, $\mathbf{l}_{\mathbf{L}} = 1$, and $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, and let $f \colon [a,b]^d \to ([u,v] \cap \mathbb{R})$ satisfy for all $x, y \in [a,b]^d$ that $|f(x) - f(y)| \leq L\|x - y\|_1$ (cf. Definition 3.3.4). Then there exists $\vartheta \in \mathbb{R}^{\mathbf{d}}$ such that $\|\vartheta\|_\infty \leq \sup_{x \in [a,b]^d} |f(x)|$ and*

$$\sup_{x \in [a,b]^d} |\mathscr{M}_{u,v}^{\vartheta,\mathbf{l}}(x) - f(x)| \leq \frac{dL(b-a)}{2} \tag{4.194}$$

(cf. Definition 4.4.1).

Proof of Lemma 4.4.11. Throughout this proof, let $\mathfrak{d} = \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, let $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_d) \in [a,b]^d$ satisfy for all $i \in \{1, 2, \ldots, d\}$ that

$$\mathbf{m}_i = \frac{a+b}{2}, \tag{4.195}$$

and let $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathbf{d}}) \in \mathbb{R}^{\mathbf{d}}$ satisfy for all $i \in \{1, 2, \ldots, \mathbf{d}\} \backslash \{\mathfrak{d}\}$ that $\vartheta_i = 0$ and $\vartheta_{\mathfrak{d}} = f(\mathbf{m})$. Note that the assumption that $\mathbf{l}_{\mathbf{L}} = 1$ and the fact that $\forall\, i \in \{1, 2, \ldots, \mathfrak{d} - 1\} \colon \vartheta_i = 0$ show that for all $x = (x_1, \ldots, x_{\mathbf{l}_{\mathbf{L}-1}}) \in \mathbb{R}^{\mathbf{l}_{\mathbf{L}-1}}$ it holds that

$$\begin{aligned}
\mathcal{A}_{1,\mathbf{l}_{\mathbf{L}-1}}^{\vartheta, \sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1}+1)}(x) &= \left[\sum_{i=1}^{\mathbf{l}_{\mathbf{L}-1}} \vartheta_{\left[\sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1}+1)\right] + i} x_i\right] + \vartheta_{\left[\sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1}+1)\right] + \mathbf{l}_{\mathbf{L}-1} + 1} \\
&= \left[\sum_{i=1}^{\mathbf{l}_{\mathbf{L}-1}} \vartheta_{\left[\sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1}+1)\right] - (\mathbf{l}_{\mathbf{L}-1} - i + 1)} x_i\right] + \vartheta_{\sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1}+1)} \\
&= \left[\sum_{i=1}^{\mathbf{l}_{\mathbf{L}-1}} \vartheta_{\mathfrak{d} - (\mathbf{l}_{\mathbf{L}-1} - i + 1)} x_i\right] + \vartheta_{\mathfrak{d}} = \vartheta_{\mathfrak{d}} = f(\mathbf{m})
\end{aligned} \tag{4.196}$$

(cf. Definition 1.1.1). Combining this with the fact that $f(\mathbf{m}) \in [u, v]$ demonstrates that for all $x \in \mathbb{R}^{\mathbf{l}_{\mathbf{L}-1}}$ it holds that

$$\begin{aligned}
\left(\mathfrak{C}_{u,v,\mathbf{l}_{\mathbf{L}}} \circ \mathcal{A}_{\mathbf{l}_{\mathbf{L}},\mathbf{l}_{\mathbf{L}-1}}^{\vartheta, \sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1}+1)}\right)(x) &= \left(\mathfrak{C}_{u,v,1} \circ \mathcal{A}_{1,\mathbf{l}_{\mathbf{L}-1}}^{\vartheta, \sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1}+1)}\right)(x) \\
&= \mathfrak{c}_{u,v}(f(\mathbf{m})) = \max\{u, \min\{f(\mathbf{m}), v\}\} \\
&= \max\{u, f(\mathbf{m})\} = f(\mathbf{m})
\end{aligned} \tag{4.197}$$

(cf. Definitions 1.2.9 and 1.2.10). This ensures for all $x \in \mathbb{R}^d$ that

$$\mathscr{M}_{u,v}^{\vartheta,\mathbf{l}}(x) = f(\mathbf{m}). \tag{4.198}$$

Furthermore, observe that (4.195) proves that for all $x \in [a, \mathbf{m}_1]$, $\mathfrak{x} \in [\mathbf{m}_1, b]$ it holds that

$$\begin{aligned}
&|\mathbf{m}_1 - x| = \mathbf{m}_1 - x = {}^{(a+b)}\!/_2 - x \leq {}^{(a+b)}\!/_2 - a = {}^{(b-a)}\!/_2 \\
\text{and} \qquad &|\mathbf{m}_1 - \mathfrak{x}| = \mathfrak{x} - \mathbf{m}_1 = \mathfrak{x} - {}^{(a+b)}\!/_2 \leq b - {}^{(a+b)}\!/_2 = {}^{(b-a)}\!/_2.
\end{aligned} \tag{4.199}$$

The assumption that $\forall\, x, y \in [a,b]^d\colon |f(x) - f(y)| \leq L\|x-y\|_1$ and (4.198) hence establish that for all $x = (x_1, \ldots, x_d) \in [a,b]^d$ it holds that

$$\begin{aligned} |\mathcal{N}^{\vartheta,\mathbf{l}}_{u,v}(x) - f(x)| &= |f(\mathbf{m}) - f(x)| \leq L\|\mathbf{m} - x\|_1 = L\sum_{i=1}^{d} |\mathbf{m}_i - x_i| \\ &= L\sum_{i=1}^{d} |\mathbf{m}_1 - x_i| \leq \sum_{i=1}^{d} \frac{L(b-a)}{2} = \frac{dL(b-a)}{2}. \end{aligned} \tag{4.200}$$

This and the fact that $\|\vartheta\|_\infty = \max_{i\in\{1,2,\ldots,\mathbf{d}\}} |\vartheta_i| = |f(\mathbf{m})| \leq \sup_{x\in[a,b]^d} |f(x)|$ imply (4.194). The proof of Lemma 4.4.11 is thus complete. □

Proposition 4.4.12. *Let $d, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $A \in (0,\infty)$, $L, a \in \mathbb{R}$, $b \in (a,\infty)$, $u \in [-\infty,\infty)$, $v \in (u,\infty]$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \ldots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, assume*

$$\mathbf{L} \geq 1 + (\lceil \log_2(A/(2d)) \rceil + 1)\mathbb{1}_{(6^d,\infty)}(A), \quad \mathbf{l}_0 = d, \quad \mathbf{l}_1 \geq A\mathbb{1}_{(6^d,\infty)}(A), \quad \mathbf{l}_{\mathbf{L}} = 1, \tag{4.201}$$

and $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1}+1)$, assume for all $i \in \{1,2,\ldots,\mathbf{L}\}\backslash\{1,\mathbf{L}\}$ that

$$\mathbf{l}_i \geq 3\lceil A/(2^i d) \rceil \mathbb{1}_{(6^d,\infty)}(A), \tag{4.202}$$

and let $f\colon [a,b]^d \to ([u,v] \cap \mathbb{R})$ satisfy for all $x, y \in [a,b]^d$ that

$$|f(x) - f(y)| \leq L\|x-y\|_1 \tag{4.203}$$

(cf. Definitions 3.3.4 and 4.2.6). Then there exists $\vartheta \in \mathbb{R}^{\mathbf{d}}$ such that $\|\vartheta\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d} |f(x)|]\}$ and

$$\sup_{x\in[a,b]^d} |\mathcal{N}^{\vartheta,\mathbf{l}}_{u,v}(x) - f(x)| \leq \frac{3dL(b-a)}{A^{1/d}} \tag{4.204}$$

(cf. Definition 4.4.1).

Proof of Proposition 4.4.12. Throughout this proof, assume without loss of generality that $A > 6^d$ (cf. Lemma 4.4.11), let $\mathfrak{Z} = \lfloor \left(\frac{A}{2d}\right)^{1/d} \rfloor \in \mathbb{Z}$. Note that the fact that for all $k \in \mathbb{N}$ it holds that $2k \leq 2(2^{k-1}) = 2^k$ shows that $3^d = 6^d/2^d \leq A/(2d)$. Therefore, we obtain that

$$2 \leq \tfrac{2}{3}\left(\tfrac{A}{2d}\right)^{1/d} \leq \left(\tfrac{A}{2d}\right)^{1/d} - 1 < \mathfrak{Z}. \tag{4.205}$$

In the next step let $r = d(b-a)/(2\mathfrak{Z}) \in (0,\infty)$, let $\delta\colon [a,b]^d \times [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that $\delta(x,y) = \|x-y\|_1$, and let $K = \max(2, \mathcal{C}^{([a,b]^d,\delta),r}) \in \mathbb{N} \cup \{\infty\}$ (cf. Definition 4.3.2). Observe that (4.205) and Lemma 4.3.4 demonstrate that

$$K = \max\{2, \mathcal{C}^{([a,b]^d,\delta),r}\} \leq \max\left\{2, \left(\lceil \tfrac{d(b-a)}{2r} \rceil\right)^d\right\} = \max\{2, (\lceil \mathfrak{Z} \rceil)^d\} = \mathfrak{Z}^d < \infty. \tag{4.206}$$

This ensures that

$$4 \leq 2dK \leq 2d\mathfrak{Z}^d \leq \tfrac{2dA}{2d} = A. \tag{4.207}$$

Combining this and the fact that $\mathbf{L} \geq 1 + (\lceil \log_2(A/(2d)) \rceil + 1)\mathbb{1}_{(6^d,\infty)}(A) = \lceil \log_2(A/(2d)) \rceil + 2$ hence proves that $\lceil \log_2(K) \rceil \leq \lceil \log_2(A/(2d)) \rceil \leq \mathbf{L}-2$. This, (4.207), the assumption that $\mathbf{l}_1 \geq A\mathbb{1}_{(6^d,\infty)}(A) = A$, and the assumption that $\forall\, i \in \{2, 3, \ldots, \mathbf{L}-1\} \colon \mathbf{l}_i \geq 3\lceil A/(2^i d) \rceil \mathbb{1}_{(6^d,\infty)}(A) = 3\lceil A/(2^i d) \rceil$ establish that for all $i \in \{2, 3, \ldots, \mathbf{L}-1\}$ it holds that

$$\mathbf{L} \geq \lceil \log_2(K) \rceil + 2, \qquad \mathbf{l}_1 \geq A \geq 2dK, \qquad \text{and} \qquad \mathbf{l}_i \geq 3\lceil \tfrac{A}{2^i d} \rceil \geq 3\lceil \tfrac{K}{2^{i-1}} \rceil. \tag{4.208}$$

Let $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_K \in [a,b]^d$ satisfy

$$\sup_{x\in[a,b]^d} \big[\inf_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\big] \leq r. \tag{4.209}$$

Note that (4.208), the assumptions that $\mathbf{l}_0 = d$, $\mathbf{l}_{\mathbf{L}} = 1$, $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, and $\forall\, x, y \in [a,b]^d \colon |f(x) - f(y)| \leq L\|x - y\|_1$, and Corollary 4.4.10 imply that there exists $\vartheta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\vartheta\|_\infty \leq \max\{1, L, \max_{k\in\{1,2,\ldots,K\}} \|\mathfrak{x}_k\|_\infty, 2\max_{k\in\{1,2,\ldots,K\}} |f(\mathfrak{x}_k)|\} \tag{4.210}$$

and

$$\begin{aligned} \sup_{x\in[a,b]^d} |\mathcal{N}^{\vartheta,\mathbf{l}}_{u,v}(x) - f(x)| &\leq 2L\big[\sup_{x\in[a,b]^d} \big(\inf_{k\in\{1,2,\ldots,K\}} \|x - \mathfrak{x}_k\|_1\big)\big] \\ &= 2L\big[\sup_{x\in[a,b]^d} \big(\inf_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\big)\big]. \end{aligned} \tag{4.211}$$

Observe that (4.210) shows that

$$\|\vartheta\|_\infty \leq \max\{1, L, |a|, |b|, 2\sup_{x\in[a,b]^d} |f(x)|\}. \tag{4.212}$$

Furthermore, note that (4.211), (4.205), (4.209), and the fact that for all $k \in \mathbb{N}$ it holds that $2k \leq 2(2^{k-1}) = 2^k$ demonstrate that

$$\begin{aligned} &\sup_{x\in[a,b]^d} |\mathcal{N}^{\vartheta,\mathbf{l}}_{u,v}(x) - f(x)| \leq 2L\big[\sup_{x\in[a,b]^d} \big(\inf_{k\in\{1,2,\ldots,K\}} \delta(x, \mathfrak{x}_k)\big)\big] \\ &\leq 2Lr = \frac{dL(b-a)}{\mathfrak{Z}} \leq \frac{dL(b-a)}{\frac{2}{3}\left(\frac{A}{2d}\right)^{1/d}} = \frac{(2d)^{1/d} 3dL(b-a)}{2A^{1/d}} \leq \frac{3dL(b-a)}{A^{1/d}}. \end{aligned} \tag{4.213}$$

Combining this with (4.212) ensures (4.204). The proof of Proposition 4.4.12 is thus complete. □

Corollary 4.4.13. *Let $d \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $L \in (0, \infty)$ and let $f \colon [a,b]^d \to \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that*

$$|f(x) - f(y)| \leq L\|x - y\|_1 \tag{4.214}$$

(cf. Definition 3.3.4). 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*

$$\mathcal{H}(\mathbf{F}) \leq \max\{0, d(\log_2(\varepsilon^{-1}) + \log_2(d) + \log_2(3L(b-a)) + 1)\}, \tag{4.215}$$

$$\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d} |f(x)|]\}, \qquad \mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \tag{4.216}$$

$$\sup_{x\in[a,b]^d} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \qquad \textit{and} \qquad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \tag{4.217}$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Corollary 4.4.13. Throughout this proof, let $\mathfrak{C} \in \mathbb{R}$ satisfy

$$\mathfrak{C} = \tfrac{9}{8}\big(3dL(b-a)\big)^{2d} + (d+22)\big(3dL(b-a)\big)^{d} + d + 11, \tag{4.218}$$

for every $\varepsilon \in (0,1]$ let $A_\varepsilon \in (0,\infty)$, $\mathbf{L}_\varepsilon \in \mathbb{N}$, $\mathbf{l}^{(\varepsilon)} = (\mathbf{l}^{(\varepsilon)}_0, \mathbf{l}^{(\varepsilon)}_1, \ldots, \mathbf{l}^{(\varepsilon)}_{\mathbf{L}_\varepsilon}) \in \mathbb{N}^{\mathbf{L}_\varepsilon+1}$ satisfy

$$A_\varepsilon = \left(\frac{3dL(b-a)}{\varepsilon}\right)^d, \qquad \mathbf{L}_\varepsilon = 1 + \big(\lceil \log_2\big(\tfrac{A_\varepsilon}{2d}\big)\rceil + 1\big)\mathbb{1}_{(6^d,\infty)}(A_\varepsilon), \tag{4.219}$$

$$\mathbf{l}^{(\varepsilon)}_0 = d, \qquad \mathbf{l}^{(\varepsilon)}_1 = \lfloor A_\varepsilon \rfloor \mathbb{1}_{(6^d,\infty)}(A_\varepsilon) + 1, \qquad \text{and} \qquad \mathbf{l}^{(\varepsilon)}_{\mathbf{L}_\varepsilon} = 1, \tag{4.220}$$

and assume for all $\varepsilon \in (0,1]$, $i \in \{2, 3, \ldots, \mathbf{L}_\varepsilon - 1\}$ that

$$\mathbf{l}^{(\varepsilon)}_i = 3\lceil \tfrac{A_\varepsilon}{2^i d}\rceil \mathbb{1}_{(6^d,\infty)}(A_\varepsilon) \tag{4.221}$$

(cf. Definition 4.2.6). Observe that the fact that for all $\varepsilon \in (0,1]$ it holds that $\mathbf{L}_\varepsilon \geq 1 + \big(\lceil \log_2\big(\frac{A_\varepsilon}{2d}\big)\rceil + 1\big)\mathbb{1}_{(6^d,\infty)}(A_\varepsilon)$, the fact that for all $\varepsilon \in (0,1]$ it holds that $\mathbf{l}^{(\varepsilon)}_0 = d$, the fact that for all $\varepsilon \in (0,1]$ it holds that $\mathbf{l}^{(\varepsilon)}_1 \geq A_\varepsilon \mathbb{1}_{(6^d,\infty)}(A_\varepsilon)$, the fact that for all $\varepsilon \in (0,1]$ it holds that $\mathbf{l}^{(\varepsilon)}_{\mathbf{L}_\varepsilon} = 1$, the fact that for all $\varepsilon \in (0,1]$, $i \in \{2, 3, \ldots, \mathbf{L}_\varepsilon - 1\}$ it holds that $\mathbf{l}^{(\varepsilon)}_i \geq 3\lceil \frac{A_\varepsilon}{2^i d}\rceil \mathbb{1}_{(6^d,\infty)}(A_\varepsilon)$, Proposition 4.4.12, and Lemma 4.4.2 prove that for all $\varepsilon \in (0,1]$ there exists $\mathbf{F}_\varepsilon \in \big(\bigtimes_{i=1}^{\mathbf{L}_\varepsilon} \big(\mathbb{R}^{\mathbf{l}^{(\varepsilon)}_i \times \mathbf{l}^{(\varepsilon)}_{i-1}} \times \mathbb{R}^{\mathbf{l}^{(\varepsilon)}_i}\big)\big) \subseteq \mathbf{N}$ which satisfies $\|\mathcal{T}(\mathbf{F}_\varepsilon)\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x\in[a,b]^d} |f(x)|]\}$ and

$$\sup_{x\in[a,b]^d} |(\mathcal{R}^{\mathbf{N}}_{\mathfrak{r}}(\mathbf{F}_\varepsilon))(x) - f(x)| \leq \frac{3dL(b-a)}{(A_\varepsilon)^{1/d}} = \varepsilon. \tag{4.222}$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6). Furthermore, observe that the fact that $d \geq 1$ establishes that for all $\varepsilon \in (0,1]$ it holds that

$$\begin{aligned} \mathcal{H}(\mathbf{F}_\varepsilon) = \mathbf{L}_\varepsilon - 1 &= \big(\lceil \log_2\big(\tfrac{A_\varepsilon}{2d}\big)\rceil + 1\big)\mathbb{1}_{(6^d,\infty)}(A_\varepsilon) \\ &= \lceil \log_2\big(\tfrac{A_\varepsilon}{d}\big)\rceil \mathbb{1}_{(6^d,\infty)}(A_\varepsilon) \leq \max\{0, \log_2(A_\varepsilon) + 1\}. \end{aligned} \tag{4.223}$$

Combining this and the fact that for all $\varepsilon \in (0,1]$ it holds that

$$\log_2(A_\varepsilon) = d\log_2\Big(\tfrac{3dL(b-a)}{\varepsilon}\Big) = d\big(\log_2(\varepsilon^{-1}) + \log_2(d) + \log_2(3L(b-a))\big) \tag{4.224}$$

implies that for all $\varepsilon \in (0,1]$ it holds that

$$\mathcal{H}(\mathbf{F}_\varepsilon) \leq \max\big\{0, d\big(\log_2(\varepsilon^{-1}) + \log_2(d) + \log_2(3L(b-a))\big) + 1\big\}. \tag{4.225}$$

Moreover, note that (4.220) and (4.221) show that for all $\varepsilon \in (0,1]$ it holds that

$$\begin{aligned}
&\mathcal{P}(\mathbf{F}_\varepsilon) = \sum_{i=1}^{\mathbf{L}_\varepsilon} \mathbf{l}_i^{(\varepsilon)}(\mathbf{l}_{i-1}^{(\varepsilon)} + 1) \\
&\leq \big(\lfloor A_\varepsilon \rfloor + 1\big)(d+1) + 3\big\lceil \tfrac{A_\varepsilon}{4d}\big\rceil\big(\lfloor A_\varepsilon \rfloor + 2\big) \\
&\quad + \max\big\{\lfloor A_\varepsilon \rfloor + 1, 3\big\lceil \tfrac{A_\varepsilon}{2^{\mathbf{L}_\varepsilon - 1}d}\big\rceil\big\} + 1 + \sum_{i=3}^{\mathbf{L}_\varepsilon - 1} 3\big\lceil \tfrac{A_\varepsilon}{2^i d}\big\rceil\big(3\big\lceil \tfrac{A_\varepsilon}{2^{i-1}d}\big\rceil + 1\big) \\
&\leq (A_\varepsilon + 1)(d+1) + 3\big(\tfrac{A_\varepsilon}{4} + 1\big)\big(A_\varepsilon + 2\big) + 3A_\varepsilon + 4 + \sum_{i=3}^{\mathbf{L}_\varepsilon - 1} 3\big(\tfrac{A_\varepsilon}{2^i} + 1\big)\big(\tfrac{3A_\varepsilon}{2^{i-1}} + 4\big).
\end{aligned} \tag{4.226}$$

In addition, observe that the fact that $\forall\, x \in (0,\infty)\colon \log_2(x) = \log_2(x/2) + 1 \leq x/2 + 1$ demonstrates that for all $\varepsilon \in (0,1]$ it holds that

$$\mathbf{L}_\varepsilon \leq 2 + \log_2(\tfrac{A_\varepsilon}{d}) \leq 3 + \tfrac{A_\varepsilon}{2d} \leq 3 + \tfrac{A_\varepsilon}{2}. \tag{4.227}$$

This ensures that for all $\varepsilon \in (0,1]$ it holds that

$$\begin{aligned}
&\sum_{i=3}^{\mathbf{L}_\varepsilon - 1} 3\big(\tfrac{A_\varepsilon}{2^i} + 1\big)\big(\tfrac{3A_\varepsilon}{2^{i-1}} + 4\big) \\
&\leq 9(A_\varepsilon)^2\Bigg[\sum_{i=3}^{\mathbf{L}_\varepsilon - 1} 2^{1-2i}\Bigg] + 12A_\varepsilon\Bigg[\sum_{i=3}^{\mathbf{L}_\varepsilon - 1} 2^{-i}\Bigg] + 9A_\varepsilon\Bigg[\sum_{i=3}^{\mathbf{L}_\varepsilon - 1} 2^{1-i}\Bigg] + 12(\mathbf{L}_\varepsilon - 3) \\
&\leq \tfrac{9(A_\varepsilon)^2}{8}\Bigg[\sum_{i=1}^{\infty} 4^{-i}\Bigg] + 3A_\varepsilon\Bigg[\sum_{i=1}^{\infty} 2^{-i}\Bigg] + \tfrac{9A_\varepsilon}{2}\Bigg[\sum_{i=1}^{\infty} 2^{-i}\Bigg] + 6A_\varepsilon \\
&= \tfrac{3}{8}(A_\varepsilon)^2 + 3A_\varepsilon + \tfrac{9}{2}A_\varepsilon + 6A_\varepsilon = \tfrac{3}{8}(A_\varepsilon)^2 + \tfrac{27}{2}A_\varepsilon.
\end{aligned} \tag{4.228}$$

This and (4.226) prove that for all $\varepsilon \in (0,1]$ it holds that

$$\begin{aligned}
\mathcal{P}(\mathbf{F}_\varepsilon) &\leq (\tfrac{3}{4} + \tfrac{3}{8})(A_\varepsilon)^2 + (d + 1 + \tfrac{9}{2} + 3 + \tfrac{27}{2})A_\varepsilon + d + 1 + 6 + 4 \\
&= \tfrac{9}{8}(A_\varepsilon)^2 + (d+22)A_\varepsilon + d + 11.
\end{aligned} \tag{4.229}$$

Combining this, (4.218), and (4.219) establishes that

$$\begin{aligned}\mathcal{P}(\mathbf{F}_\varepsilon) &\leq \tfrac{9}{8}\big(3dL(b-a)\big)^{2d}\varepsilon^{-2d} + (d+22)\big(3dL(b-a)\big)^{d}\varepsilon^{-d} + d + 11 \\ &\leq \Big[\tfrac{9}{8}\big(3dL(b-a)\big)^{2d} + (d+22)\big(3dL(b-a)\big)^{d} + d + 11\Big]\varepsilon^{-2d} = \mathfrak{C}\varepsilon^{-2d}.\end{aligned} \tag{4.230}$$

Combining this with (4.222) and (4.225) proves (4.215), (4.216), and (4.217). The proof of Corollary 4.4.13 is thus complete. □

Remark 4.4.14 (High-dimensional ANN approximation results). Corollary 4.4.13 above is a multi-dimensional ANN approximation result in the sense that the input dimension $d \in \mathbb{N}$ of the domain of definition $[a,b]^d$ of the considered target function f that we intend to approximate can be any natural number. However, we note that Corollary 4.4.13 does not provide a useful contribution in the case when the dimension d is large, say $d \geq 5$, as Corollary 4.4.13 does not provide any information on how the constant $\mathfrak{C}$ in (4.217) grows in d and as the dimension d appears in the exponent of the reciprocal ε^{-1} of the prescribed approximation accuracy ε in the bound for the number of ANN parameters in (4.217).

In the literature there are also a number of suitable high-dimensional ANN approximation results which assure that the constant in the parameter bound grows at most polynomially in the dimension d and which assure that the exponent of the reciprocal ε^{-1} of the prescribed approximation accuracy ε in the ANN parameter bound is completely independent of the dimension d. Such results do have the potential to provide a useful practical conclusion for ANN approximations even when the dimension d is large. We refer, for example, to [14, 15, 28, 72, 126, 166] and the references therein for such high-dimensional ANN approximation results in the context of general classes of target functions and we refer, for instance, to [3, 29, 35, 128, 133, 167–169, 183, 185, 213, 217, 238, 269, 367] and the references therein for such high-dimensional ANN approximation results where the target functions are solutions of PDEs (cf. also Section 18.4 below).

Remark 4.4.15 (Infinite-dimensional ANN approximation results). In the literature there are now also results where the target function that we intend to approximate is defined on an infinite-dimensional vector space and where the dimension of the domain of definition of the target function is thus infinity (see, for example, [32, 69, 70, 210, 265, 377] and the references therein). This perspective seems to be very reasonable as in many applications, input data, such as images and videos, that should be processed through the target function are more naturally represented by elements of infinite-dimensional spaces instead of elements of finite-dimensional spaces.

Part III

Optimization

Chapter 5

Optimization through gradient flow (GF) trajectories

In Chapters 6 and 7 below we study deterministic and stochastic GD-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 7 can be viewed as suitable Monte Carlo approximations of the deterministic GD-type optimization methods in Chapter 6 and the deterministic GD-type optimization methods in Chapter 6 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.

Further investigations on optimization through GF ODEs can, for example, be found in [2, 45, 131, 224, 233, 234, 268] and the references therein.

5.1 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 gradient based optimization methods arise within deep supervised learning algorithms. To do this, we now recall the deep supervised learning framework from the introduction.

Specifically, let $d, M \in \mathbb{N}$, $\mathcal{E} \in C(\mathbb{R}^d, \mathbb{R})$, $x_1, x_2, \ldots, x_{M+1} \in \mathbb{R}^d$, $y_1, y_2, \ldots, y_M \in \mathbb{R}$ satisfy for all $m \in \{1, 2, \ldots, M\}$ that

$$y_m = \mathcal{E}(x_m) \tag{5.1}$$

and let $\mathfrak{L}\colon C(\mathbb{R}^d, \mathbb{R}) \to [0, \infty)$ satisfy for all $\phi \in C(\mathbb{R}^d, \mathbb{R})$ that

$$\mathfrak{L}(\phi) = \frac{1}{M}\left[\sum_{m=1}^{M} |\phi(x_m) - y_m|^2\right]. \tag{5.2}$$

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 $\mathcal{E}\colon \mathbb{R}^d \to \mathbb{R}$ as an unknown function which relates input and output data through (5.1), we think of $x_1, x_2, \ldots, x_{M+1} \in \mathbb{R}^d$ as the available known input data, we think of $y_1, y_2, \ldots, y_M \in \mathbb{R}$ as the available known output data, and we have that the function $\mathfrak{L}\colon C(\mathbb{R}^d, \mathbb{R}) \to [0, \infty)$ in (5.2) 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

$$\mathfrak{L}(\mathcal{E}) = 0 \tag{5.3}$$

and we are trying to approximate the function $\mathcal{E}$ by computing an approximate minimizer of the function $\mathfrak{L}\colon C(\mathbb{R}^d, \mathbb{R}) \to [0, \infty)$. In order to make this optimization problem amenable to numerical computations, we consider a spatially discretized version of the optimization problem associated to (5.2) by employing parametrizations of ANNs (cf. (7) in the introduction).

More formally, let $a\colon \mathbb{R} \to \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \ldots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + \left[\sum_{k=2}^{h} l_k(l_{k-1}+1)\right] + l_h + 1$, and consider the parametrization function

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathcal{N}^{\theta,d}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \ldots, \mathfrak{M}_{a,l_h}, \mathrm{id}_{\mathbb{R}}} \in C(\mathbb{R}^d, \mathbb{R}) \tag{5.4}$$

(cf. Definitions 1.1.3 and 1.2.1). Note that h is the number of hidden layers of the ANNs in (5.4), note for every $i \in \{1, 2, \ldots, h\}$ that $l_i \in \mathbb{N}$ is the number of neurons in the i-th hidden layer of the ANNs in (5.4), and note that $\mathfrak{d}$ is the number of real parameters used to describe the ANNs in (5.4). Observe that for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ we have that the function

$$\mathbb{R}^d \ni x \mapsto \mathcal{N}^{\theta,d}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \ldots, \mathfrak{M}_{a,l_h}, \mathrm{id}_{\mathbb{R}}} \in \mathbb{R} \tag{5.5}$$

in (5.4) is nothing else than the realization function associated to a *fully-connected feedforward* ANN where before each hidden layer a multi-dimensional version of the activation function $a\colon \mathbb{R} \to \mathbb{R}$ is applied. We restrict ourselves in this section to a differentiable activation function as this differentiability property allows us to consider gradients (cf. (5.7), (5.8), and Section 5.3.2 below for details).

We now discretize the optimization problem in (5.2) as the problem of computing approximate minimizers of the function $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty)$ which satisfies for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \left|\left(\mathcal{N}^{\theta,d}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \ldots, \mathfrak{M}_{a,l_h}, \mathrm{id}_{\mathbb{R}}}\right)(x_m) - y_m\right|^2\right] \tag{5.6}$$

and this resulting optimization problem is now accessible to numerical computations. Specifically, deep learning algorithms solve optimization problems of the type (5.6) by means of *gradient based optimization methods*. Loosely speaking, gradient based optimization methods aim to minimize the considered objective function (such as (5.6) 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 and we now sketch the GD optimization method applied to (5.6). Let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, and let $\theta = (\theta_n)_{n\in\mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\theta_0 = \xi \qquad \text{and} \qquad \theta_n = \theta_{n-1} - \gamma_n (\nabla \mathscr{L})(\theta_{n-1}). \tag{5.7}$$

The process $(\theta_n)_{n\in\mathbb{N}_0}$ is the GD process for the minimization problem associated to (5.6) with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ (see Definition 6.1.1 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 (5.6) such solutions of GF ODEs can be described as follows. Let $\Theta = (\Theta_t)_{t\in[0,\infty)} \colon [0,\infty) \to \mathbb{R}^{\mathfrak{d}}$ be a continuously differentiable function which satisfies for all $t \in [0,\infty)$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \dot{\Theta}_t = \tfrac{\partial}{\partial t}\Theta_t = -(\nabla \mathscr{L})(\Theta_t). \tag{5.8}$$

The process $(\Theta_t)_{t\in[0,\infty)}$ is the solution of the GF ODE corresponding to the minimization problem associated to (5.6) with initial value ξ.

In Chapter 6 below we introduce and study deterministic GD-type optimization methods such as the GD optimization method in (5.7). 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 (5.8) above. In deep learning algorithms usually not GD-type optimization methods but stochastic variants of GD-type optimization methods are employed to solve optimization problems of the form (5.6). Such SGD-type optimization methods can be viewed as suitable Monte Carlo approximations of deterministic GD-type methods and in Chapter 7 below we treat such SGD-type optimization methods.

5.2 Basics for GFs

5.2.1 GF ordinary differential equations (ODEs)

Definition 5.2.1 (GF trajectories)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function, and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a $\mathcal{B}(\mathbb{R}^{\mathfrak{d}})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$-measurable function which satisfies for all $U \in \{V \subseteq \mathbb{R}^{\mathfrak{d}} \colon V \text{ is open}\}$, $\theta \in U$ with $\mathscr{L}|_U \in C^1(U, \mathbb{R}^{\mathfrak{d}})$ that*

$$\mathscr{G}(\theta) = (\nabla \mathscr{L})(\theta). \tag{5.9}$$

Then we say that Θ is a GF trajectory for the objective function $\mathscr{L}$ with generalized gradient $\mathscr{G}$ and initial value ξ (we say that Θ is a GF trajectory for the objective function $\mathscr{L}$ with initial value ξ, we say that Θ is a solution of the GF ODE for the objective function $\mathscr{L}$ with generalized gradient $\mathscr{G}$ and initial value ξ, we say that Θ is a solution of the GF ODE for the objective function $\mathscr{L}$ with initial value ξ) if and only if it holds that $\Theta \colon [0, \infty) \to \mathbb{R}^{\mathfrak{d}}$ is a continuous function from $[0, \infty)$ to $\mathbb{R}^{\mathfrak{d}}$ which satisfies for all $t \in [0, \infty)$ that $\int_0^t \|\mathscr{G}(\Theta_s)\|_2 \, \mathrm{d}s < \infty$ and

$$\Theta_t = \xi - \int_0^t \mathscr{G}(\Theta_s) \, \mathrm{d}s \tag{5.10}$$

(cf. Definition 3.3.4).

5.2.2 Direction of negative gradients

Lemma 5.2.2. *Let $\mathfrak{d} \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $r \in (0, \infty)$ and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{G}(v) = \lim_{h \to 0} \left(\frac{\mathscr{L}(\theta + hv) - \mathscr{L}(\theta)}{h} \right) = [\mathscr{L}'(\theta)](v). \tag{5.11}$$

Then

(i) it holds that

$$\sup_{v \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w\|_2 = r\}} \mathscr{G}(v) = r \|(\nabla \mathscr{L})(\theta)\|_2 = \begin{cases} 0 & : (\nabla \mathscr{L})(\theta) = 0 \\ \mathscr{G}\left(\frac{r(\nabla \mathscr{L})(\theta)}{\|(\nabla \mathscr{L})(\theta)\|_2} \right) & : (\nabla \mathscr{L})(\theta) \neq 0 \end{cases} \tag{5.12}$$

and

(ii) it holds that

$$\inf_{v\in\{w\in\mathbb{R}^{\mathfrak{d}}\colon \|w\|_2=r\}} \mathscr{G}(v) = -r\|(\nabla\mathscr{L})(\theta)\|_2 = \begin{cases} 0 & \colon (\nabla\mathscr{L})(\theta) = 0 \\ \mathscr{G}\big(\frac{-r(\nabla\mathscr{L})(\theta)}{\|(\nabla\mathscr{L})(\theta)\|_2}\big) & \colon (\nabla\mathscr{L})(\theta) \neq 0 \end{cases} \tag{5.13}$$

(cf. Definition 3.3.4).

Proof of Lemma 5.2.2. Note that (5.11) implies that for all $v \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathscr{G}(v) = \langle(\nabla\mathscr{L})(\theta), v\rangle \tag{5.14}$$

(cf. Definition 1.4.7). The Cauchy–Schwarz inequality therefore ensures that for all $v \in \mathbb{R}^{\mathfrak{d}}$ with $\|v\|_2 = r$ it holds that

$$\begin{aligned} &-r\|(\nabla\mathscr{L})(\theta)\|_2 = -\|(\nabla\mathscr{L})(\theta)\|_2\|v\|_2 \leq -\langle-(\nabla\mathscr{L})(\theta), v\rangle \\ &= \mathscr{G}(v) \leq \|(\nabla\mathscr{L})(\theta)\|_2\|v\|_2 = r\|(\nabla\mathscr{L})(\theta)\|_2 \end{aligned} \tag{5.15}$$

(cf. Definition 3.3.4). Furthermore, note that (5.14) shows that for all $c \in \mathbb{R}$ it holds that

$$\mathscr{G}(c(\nabla\mathscr{L})(\theta)) = \langle(\nabla\mathscr{L})(\theta), c(\nabla\mathscr{L})(\theta)\rangle = c\|(\nabla\mathscr{L})(\theta)\|_2^2. \tag{5.16}$$

Combining this and (5.15) proves item (i) and item (ii). The proof of Lemma 5.2.2 is thus complete. □

Lemma 5.2.3. *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and assume for all $t \in [0,\infty)$ that $\Theta_t = \Theta_0 - \int_0^t (\nabla\mathscr{L})(\Theta_s)\,\mathrm{d}s$. Then*

(i) it holds that $\Theta \in C^1([0,\infty), \mathbb{R}^{\mathfrak{d}})$,

(ii) it holds for all $t \in [0,\infty)$ that $\dot{\Theta}_t = -(\nabla\mathscr{L})(\Theta_t)$, and

(iii) it holds for all $t \in [0,\infty)$ that

$$\mathscr{L}(\Theta_t) = \mathscr{L}(\Theta_0) - \int_0^t \|(\nabla\mathscr{L})(\Theta_s)\|_2^2\,\mathrm{d}s \tag{5.17}$$

(cf. Definition 3.3.4).

Proof of Lemma 5.2.3. Observe that the fundamental theorem of calculus implies item (i) and item (ii). Combining item (ii) with the fundamental theorem of calculus and the chain rule ensures that for all $t \in [0,\infty)$ it holds that

$$\mathscr{L}(\Theta_t) = \mathscr{L}(\Theta_0) + \int_0^t \langle(\nabla\mathscr{L})(\Theta_s), \dot{\Theta}_s\rangle\,\mathrm{d}s = \mathscr{L}(\Theta_0) - \int_0^t \|(\nabla\mathscr{L})(\Theta_s)\|_2^2\,\mathrm{d}s \tag{5.18}$$

(cf. Definitions 1.4.7 and 3.3.4). This establishes item (iii). The proof of Lemma 5.2.3 is thus complete. □

Corollary 5.2.4 (Illustration for the negative GF)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *and assume for all* $t \in [0,\infty)$ *that* $\Theta(t) = \Theta(0) - \int_0^t (\nabla \mathscr{L})(\Theta(s))\,\mathrm{d}s$. *Then*

(i) it holds that $\Theta \in C^1([0,\infty), \mathbb{R}^{\mathfrak{d}})$,

(ii) it holds for all $t \in (0,\infty)$ *that*

$$(\mathscr{L} \circ \Theta)'(t) = -\|(\nabla \mathscr{L})(\Theta(t))\|_2^2, \tag{5.19}$$

and

(iii) it holds for all $\Xi \in C^1([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\tau \in (0,\infty)$ *with* $\Xi(\tau) = \Theta(\tau)$ *and* $\|\Xi'(\tau)\|_2 = \|\Theta'(\tau)\|_2$ *that*

$$(\mathscr{L} \circ \Theta)'(\tau) \leq (\mathscr{L} \circ \Xi)'(\tau) \tag{5.20}$$

(cf. Definition 3.3.4).

Proof of Corollary 5.2.4. Note that Lemma 5.2.3 and the fundamental theorem of calculus imply items (i) and (ii). Observe that Lemma 5.2.2 shows for all $\Xi \in C^1([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $t \in (0,\infty)$ it holds that

$$\begin{aligned}(\mathscr{L} \circ \Xi)'(t) &= [\mathscr{L}'(\Xi(t))](\Xi'(t)) \\ &\geq \inf_{v \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w\|_2 = \|\Xi'(t)\|_2\}} [\mathscr{L}'(\Xi(t))](v) \\ &= -\|\Xi'(t)\|_2 \|(\nabla \mathscr{L})(\Xi(t))\|_2\end{aligned} \tag{5.21}$$

(cf. Definition 3.3.4). Lemma 5.2.3 hence ensures that for all $\Xi \in C^1([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\tau \in (0,\infty)$ with $\Xi(\tau) = \Theta(\tau)$ and $\|\Xi'(\tau)\|_2 = \|\Theta'(\tau)\|_2$ it holds that

$$\begin{aligned}(\mathscr{L} \circ \Xi)'(\tau) &\geq -\|\Xi'(\tau)\|_2 \|(\nabla \mathscr{L})(\Xi(\tau))\|_2 \geq -\|\Theta'(\tau)\|_2 \|(\nabla \mathscr{L})(\Theta(\tau))\|_2 \\ &= -\|(\nabla \mathscr{L})(\Theta(\tau))\|_2^2 = (\mathscr{L} \circ \Theta)'(\tau).\end{aligned} \tag{5.22}$$

This establishes item (iii). The proof of Corollary 5.2.4 is thus complete. □

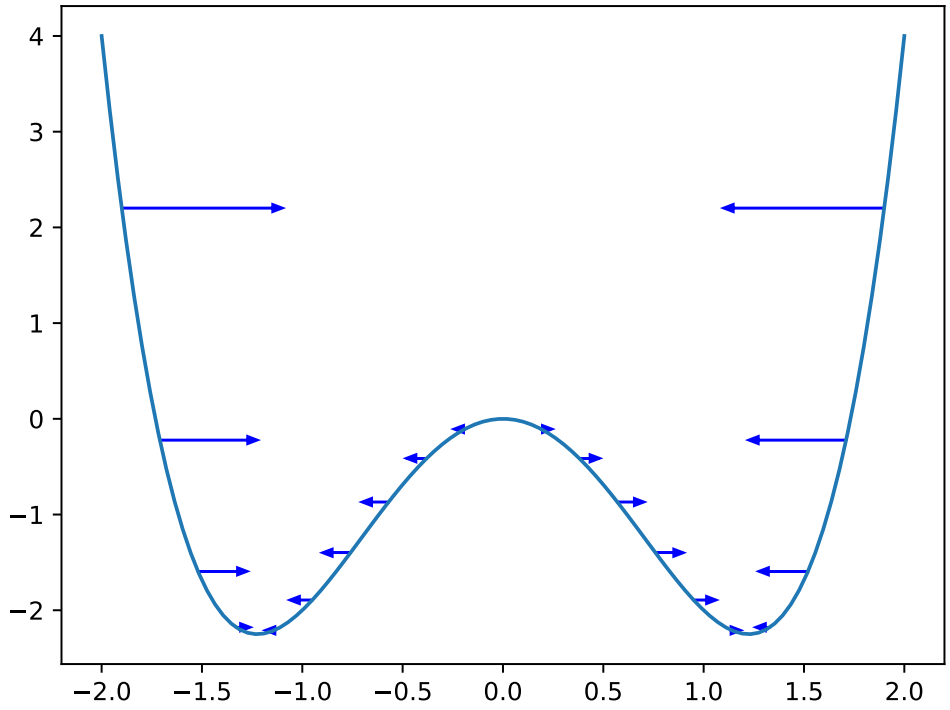

Figure 5.1 (plots/gradient_plot1.pdf): Illustration of negative gradients in a one-dimensional example. The plot shows the graph of the function $[-2, 2] \ni x \mapsto x^4 - 3x^2 \in \mathbb{R}$ with the value of the negative gradient, scaled by $\frac{1}{20}$, indicated by horizontal arrows at several points. The PYTHON code used to produce this plot is given in Source code 5.1.

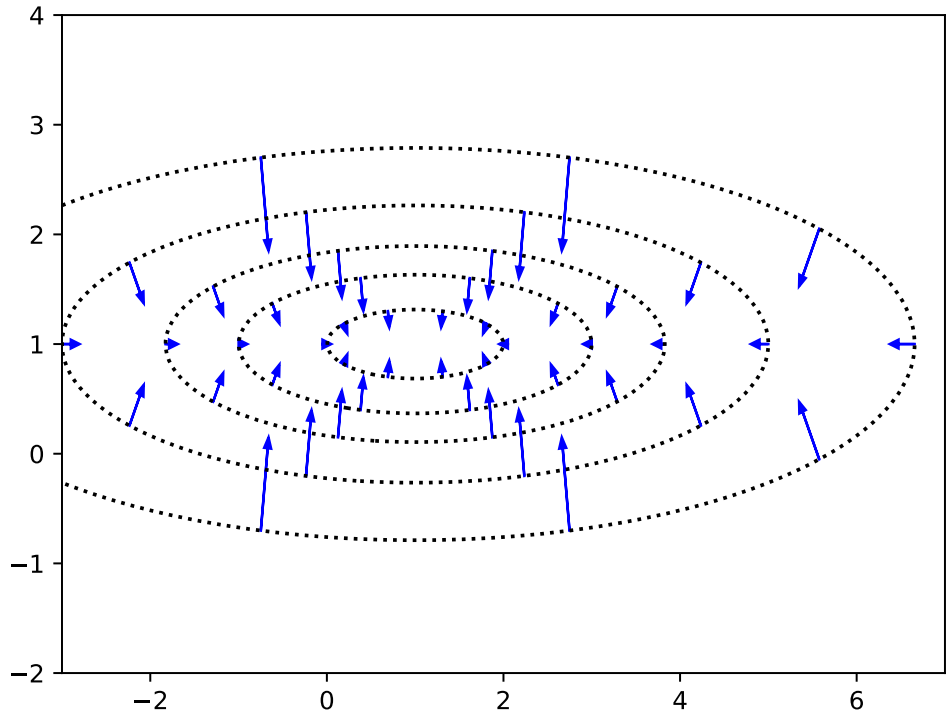

Figure 5.2 (plots/gradient_plot2.pdf): Illustration of negative gradients in a two-dimensional example. The plot shows contour lines of the function $\mathbb{R}^2 \ni (x, y) \mapsto \frac{1}{2}|x-1|^2 + 5|y-1|^2 \in \mathbb{R}$ with arrows indicating the direction and magnitude, scaled by $\frac{1}{20}$, of the negative gradient at several points along these contour lines. The PYTHON code used to produce this plot is given in Source code 5.2.

```
import numpy as np
import matplotlib.pyplot as plt

def f(x):
  return x**4 - 3 * x**2

def nabla_f(x):
  return 4 * x**3 - 6 * x

plt.figure()

# Plot graph of f
x = np.linspace(-2,2,100)
```

```
plt.plot(x,f(x))

# Plot arrows
for x in np.linspace(-1.9,1.9,21):
  d = nabla_f(x)
  plt.arrow(x, f(x), -.05 * d, 0,
    length_includes_head=True, head_width=0.08,
    head_length=0.05, color='b')

plt.savefig("../plots/gradient_plot1.pdf")
```

Source code 5.1 (code/gradient_plot1.py): PYTHON code used to create Figure 5.1

```
import numpy as np
import matplotlib.pyplot as plt

K = [1., 10.]
vartheta = np.array([1., 1.])

def f(x, y):
    result =  K[0] / 2. * np.abs(x - vartheta[0])**2 \
    + K[1] / 2. * np.abs(y - vartheta[1])**2
    return result

def nabla_f(x):
    return K * (x - vartheta)

plt.figure()

# Plot contour lines of f
x = np.linspace(-3., 7., 100)
y = np.linspace(-2., 4., 100)
X, Y = np.meshgrid(x, y)
Z = f(X, Y)
cp = plt.contour(X, Y, Z, colors="black",
                 levels = [0.5,2,4,8,16],
                 linestyles=":")

# Plot arrows along contour lines
for l in [0.5,2,4,8,16]:
  for d in np.linspace(0, 2.*np.pi, 10, endpoint=False):
    x = np.cos(d) / ((K[0] / (2*l))**.5) + vartheta[0]
    y = np.sin(d) / ((K[1] / (2*l))**.5) + vartheta[1]
    grad = nabla_f(np.array([x,y]))
    plt.arrow(x, y, -.05 * grad[0], -.05 * grad[1],
      length_includes_head=True, head_width=.08,
      head_length=.1, color='b')

plt.savefig("../plots/gradient_plot2.pdf")
```

Source code 5.2 (code/gradient_plot2.py): PYTHON code used to create Figure 5.2

5.3 Regularity properties for ANNs

5.3.1 On the differentiability of compositions of parametric functions

Lemma 5.3.1. *Let $\mathfrak{d}_1, \mathfrak{d}_2, l_1, l_2 \in \mathbb{N}$, let $A_1 \colon \mathbb{R}^{l_1} \to \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ and $A_2 \colon \mathbb{R}^{l_2} \to \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ satisfy for all $x_1 \in \mathbb{R}^{l_1}$, $x_2 \in \mathbb{R}^{l_2}$ that $A_1(x_1) = (x_1, 0)$ and $A_2(x_2) = (0, x_2)$, for every $k \in \{1, 2\}$ let $B_k \colon \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} \to \mathbb{R}^{l_k}$ satisfy for all $x_1 \in \mathbb{R}^{l_1}$, $x_2 \in \mathbb{R}^{l_2}$ that $B_k(x_1, x_2) = x_k$, for every $k \in \{1, 2\}$ let $F_k \colon \mathbb{R}^{\mathfrak{d}_k} \to \mathbb{R}^{l_k}$ be differentiable, and let $f \colon \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \to \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ satisfy for all $x_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $x_2 \in \mathbb{R}^{\mathfrak{d}_2}$ that*

$$f(x_1, x_2) = (F_1(x_1), F_2(x_2)). \tag{5.23}$$

Then

(i) it holds that $f = A_1 \circ F_1 \circ B_1 + A_2 \circ F_2 \circ B_2$ and

(ii) it holds that f is differentiable.

Proof of Lemma 5.3.1. Note that (5.23) implies that for all $x_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $x_2 \in \mathbb{R}^{\mathfrak{d}_2}$ it holds that

$$\begin{aligned}(A_1 \circ F_1 \circ B_1 + A_2 \circ F_2 \circ B_2)(x_1, x_2) &= (A_1 \circ F_1)(x_1) + (A_2 \circ F_2)(x_2) \\ &= (F_1(x_1), 0) + (0, F_2(x_2)) \\ &= (F_1(x_1), F_2(x_2)).\end{aligned} \tag{5.24}$$

Combining this and the fact that A_1, A_2, F_1, F_2, B_1, and B_2 are differentiable with the chain rule establishes that f is differentiable. The proof of Lemma 5.3.1 is thus complete. □

Lemma 5.3.2. *Let $\mathfrak{d}_1, \mathfrak{d}_2, l_0, l_1, l_2 \in \mathbb{N}$, let $A \colon \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_0} \to \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}_1 + l_0}$ and $B \colon \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}_1 + l_0} \to \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_1}$ satisfy for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, $x \in \mathbb{R}^{l_0}$ that*

$$A(\theta_1, \theta_2, x) = (\theta_2, (\theta_1, x)) \qquad \text{and} \qquad B(\theta_2, (\theta_1, x)) = (\theta_2, F_1(\theta_1, x)), \tag{5.25}$$

for every $k \in \{1, 2\}$ let $F_k \colon \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \to \mathbb{R}^{l_k}$ be differentiable, and let $f \colon \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_0} \to \mathbb{R}^{l_2}$ satisfy for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, $x \in \mathbb{R}^{l_0}$ that

$$f(\theta_1, \theta_2, x) = \big(F_2(\theta_2, \cdot) \circ F_1(\theta_1, \cdot)\big)(x). \tag{5.26}$$

Then

(i) it holds that $f = F_2 \circ B \circ A$ *and*

(ii) it holds that f *is differentiable.*

Proof of Lemma 5.3.2. Observe that (5.25) and (5.26) show that for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, $x \in \mathbb{R}^{l_0}$ it holds that

$$f(\theta_1, \theta_2, x) = F_2(\theta_2, F_1(\theta_1, x)) = F_2(B(\theta_2, (\theta_1, x))) = F_2(B(A(\theta_1, \theta_2, x))). \tag{5.27}$$

Note that Lemma 5.3.1 (applied with $\mathfrak{d}_1 \curvearrowleft \mathfrak{d}_2$, $\mathfrak{d}_2 \curvearrowleft \mathfrak{d}_1 + l_1$, $l_1 \curvearrowleft \mathfrak{d}_2$, $l_2 \curvearrowleft l_1$, $F_1 \curvearrowleft (\mathbb{R}^{\mathfrak{d}_2} \ni \theta_2 \mapsto \theta_2 \in \mathbb{R}^{\mathfrak{d}_2})$, $F_2 \curvearrowleft (\mathbb{R}^{\mathfrak{d}_1 + l_1} \ni (\theta_1, x) \mapsto F_1(\theta_1, x) \in \mathbb{R}^{l_1})$ in the notation of Lemma 5.3.1) implies that B is differentiable. Combining this, the fact that A is differentiable, the fact that F_2 is differentiable, and (5.27) with the chain rule assures that f is differentiable. The proof of Lemma 5.3.2 is thus complete. □

5.3.2 On the differentiability of realizations of ANNs

Lemma 5.3.3 (Differentiability of realization functions of ANNs)**.** *Let* $L \in \mathbb{N}$, $l_0, l_1, \ldots, l_L \in \mathbb{N}$, *for every* $k \in \{1, 2, \ldots, L\}$ *let* $\mathfrak{d}_k = l_k(l_{k-1} + 1)$, *for every* $k \in \{1, 2, \ldots, L\}$ *let* $\Psi_k \colon \mathbb{R}^{l_k} \to \mathbb{R}^{l_k}$ *be differentiable, and for every* $k \in \{1, 2, \ldots, L\}$ *let* $F_k \colon \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \to \mathbb{R}^{l_k}$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}_k}$, $x \in \mathbb{R}^{l_{k-1}}$ *that*

$$F_k(\theta, x) = \Psi_k\big(\mathcal{A}^{\theta,0}_{l_k, l_{k-1}}(x)\big) \tag{5.28}$$

(cf. Definition 1.1.1). Then

(i) it holds for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, …, $\theta_L \in \mathbb{R}^{\mathfrak{d}_L}$, $x \in \mathbb{R}^{l_0}$ *that*

$$\big(\mathcal{N}^{(\theta_1, \theta_2, \ldots, \theta_L), l_0}_{\Psi_1, \Psi_2, \ldots, \Psi_L}\big)(x) = (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \ldots \circ F_1(\theta_1, \cdot))(x) \tag{5.29}$$

and

(ii) it holds that

$$\mathbb{R}^{\mathfrak{d}_1 + \mathfrak{d}_2 + \ldots + \mathfrak{d}_L} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto \big(\mathcal{N}^{\theta, l_0}_{\Psi_1, \Psi_2, \ldots, \Psi_L}\big)(x) \in \mathbb{R}^{l_L} \tag{5.30}$$

is differentiable

(cf. Definition 1.1.3).

Proof of Lemma 5.3.3. Observe that (1.1) shows that for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, …,

$\theta_L \in \mathbb{R}^{\mathfrak{d}_L}$, $k \in \{1, 2, \ldots, L\}$ it holds that

$$\mathcal{A}^{(\theta_1,\theta_2,\ldots,\theta_L),\sum_{j=1}^{k-1}\mathfrak{d}_j}_{l_k,l_{k-1}} = \mathcal{A}^{\theta_k,0}_{l_k,l_{k-1}}. \tag{5.31}$$

Therefore, we obtain that for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, …, $\theta_L \in \mathbb{R}^{\mathfrak{d}_L}$, $k \in \{1, 2, \ldots, L\}$ it holds that

$$F_k(\theta_k, x) = \big(\Psi_k \circ \mathcal{A}^{(\theta_1,\theta_2,\ldots,\theta_L),\sum_{j=1}^{k-1}\mathfrak{d}_j}_{l_k,l_{k-1}}\big)(x). \tag{5.32}$$

Combining this with (1.5) establishes item (i). Note that the assumption that for all $k \in \{1, 2, \ldots, L\}$ it holds that Ψ_k is differentiable, the fact that for all $m, n \in \mathbb{N}$, $\theta \in \mathbb{R}^{m(n+1)}$ it holds that $\mathbb{R}^{m(n+1)} \times \mathbb{R}^n \ni (\theta, x) \mapsto \mathcal{A}^{\theta,0}_{m,n}(x) \in \mathbb{R}^m$ is differentiable, and the chain rule ensure that for all $k \in \{1, 2, \ldots, L\}$ it holds that F_k is differentiable. Lemma 5.3.2 and induction hence prove that

$$\begin{aligned}\mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \ldots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_0} \ni (\theta_1, \theta_2, \ldots, \theta_L, x)& \\ \mapsto (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \ldots \circ F_1(\theta_1, \cdot))(x) \in \mathbb{R}^{l_L}&\end{aligned} \tag{5.33}$$

is differentiable. This and item (i) prove item (ii). The proof of Lemma 5.3.3 is thus complete. □

Lemma 5.3.4 (Differentiability of the empirical risk function)**.** *Let $L, \mathfrak{d} \in \mathbb{N}\backslash\{1\}$, $M, l_0, l_1, \ldots, l_L \in \mathbb{N}$, $x_1, x_2, \ldots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \ldots, y_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^{L} l_k(l_{k-1}+1)$, for every $k \in \{1, 2, \ldots, L\}$ let $\Psi_k \colon \mathbb{R}^{l_k} \to \mathbb{R}^{l_k}$ be differentiable, and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbf{L}\big(\big(\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L}\big)(x_m), y_m\big)\right] \tag{5.34}$$

(cf. Definition 1.1.3). Then $\mathscr{L}$ is differentiable.

Proof of Lemma 5.3.4. Observe that Lemma 5.3.3 and Lemma 5.3.1 (applied with $\mathfrak{d}_1 \curvearrowleft \mathfrak{d} + l_0$, $\mathfrak{d}_2 \curvearrowleft l_L$, $l_1 \curvearrowleft l_L$, $l_2 \curvearrowleft l_L$, $F_1 \curvearrowleft (\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto \big(\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L}\big)(x) \in \mathbb{R}^{l_L})$, $F_2 \curvearrowleft \mathrm{id}_{\mathbb{R}^{l_L}}$ in the notation of Lemma 5.3.1) imply that

$$\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{l_0} \times \mathbb{R}^{l_L} \ni (\theta, x, y) \mapsto \big(\big(\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L}\big)(x), y\big) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \tag{5.35}$$

is differentiable. The assumption that $\mathbf{L}$ is differentiable and the chain rule hence demonstrate that for all $x \in \mathbb{R}^{l_0}$, $y \in \mathbb{R}^{l_L}$ it holds that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathbf{L}\big(\big(\mathcal{N}^{\theta,l_0}_{\Psi_1,\Psi_2,\ldots,\Psi_L}\big)(x_m), y_m\big) \in \mathbb{R} \tag{5.36}$$

is differentiable. This ensures that $\mathscr{L}$ is differentiable. The proof of Lemma 5.3.4 is thus complete. □

Lemma 5.3.5. *Let $a\colon \mathbb{R} \to \mathbb{R}$ be differentiable and let $d \in \mathbb{N}$. Then $\mathfrak{M}_{a,d}$ is differentiable (cf. Definition 1.2.1).*

Proof of Lemma 5.3.5. Note that the assumption that a is differentiable, Lemma 5.3.1, and induction establish that for all $m \in \mathbb{N}$ it holds that $\mathfrak{M}_{a,m}$ is differentiable. The proof of Lemma 5.3.5 is thus complete. □

Corollary 5.3.6. *Let $L, \mathfrak{d} \in \mathbb{N}\backslash\{1\}$, $M, l_0, l_1, \ldots, l_L \in \mathbb{N}$, $x_1, x_2, \ldots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \ldots, y_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^{L} l_k(l_{k-1}+1)$, let $a\colon \mathbb{R} \to \mathbb{R}$ and $\mathbf{L}\colon \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \to \mathbb{R}$ be differentiable, and let $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbf{L}\big(\big(\mathcal{N}^{\theta,l_0}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},\mathrm{id}_{\mathbb{R}^{l_L}}}\big)(x_m), y_m\big)\right] \tag{5.37}$$

(cf. Definitions 1.1.3 and 1.2.1). Then $\mathscr{L}$ is differentiable.

Proof of Corollary 5.3.6. Observe that Lemma 5.3.5, and Lemma 5.3.4 prove that $\mathscr{L}$ is differentiable. The proof of Corollary 5.3.6 is thus complete. □

Corollary 5.3.7. *Let $L, \mathfrak{d} \in \mathbb{N}\backslash\{1\}$, $M, l_0, l_1, \ldots, l_L \in \mathbb{N}$, $x_1, x_2, \ldots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \ldots, y_M \in (0,\infty)^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^{L} l_k(l_{k-1}+1)$, let A be the l_L-dimensional softmax activation function, let $a\colon \mathbb{R} \to \mathbb{R}$ and $\mathbf{L}\colon (0,\infty)^{l_L} \times (0,\infty)^{l_L} \to \mathbb{R}$ be differentiable, and let $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbf{L}\big(\big(\mathcal{N}^{\theta,l_0}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},A}\big)(x_m), y_m\big)\right] \tag{5.38}$$

(cf. Definitions 1.1.3, 1.2.1, and 1.2.43 and Lemma 1.2.44). Then $\mathscr{L}$ is differentiable.

Proof of Corollary 5.3.7. Note that Lemma 5.3.5, the fact that A is differentiable, and Lemma 5.3.4 show that $\mathscr{L}$ is differentiable. The proof of Corollary 5.3.7 is thus complete. □

5.4 Loss functions

5.4.1 Absolute error loss

Definition 5.4.1. *Let $d \in \mathbb{N}$ and let $\|\!|\cdot|\!\|\colon \mathbb{R}^d \to [0,\infty)$ be a norm. Then we say that $\mathbf{L}$ is the ℓ^1-error loss function based on $\|\!|\cdot|\!\|$ (we say that $\mathbf{L}$ is the absolute error loss function based on $\|\!|\cdot|\!\|$) if and only if it holds that $\mathbf{L}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 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*

$$\mathbf{L}(x, y) = \|\!|x - y|\!\|. \tag{5.39}$$

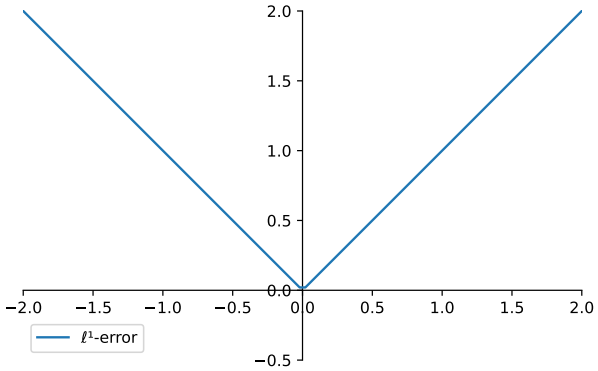

Figure 5.3 (plots/l1loss.pdf): A plot of the function $\mathbb{R} \ni x \mapsto \mathbf{L}(x, 0) \in [0, \infty)$ where $\mathbf{L}$ is the ℓ^1-error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$ (cf. Definition 5.4.1).

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-2,2), (-.5,2))

x = np.linspace(-2, 2, 100)

mae_loss = tf.keras.losses.MeanAbsoluteError(
    reduction=tf.keras.losses.Reduction.NONE)
zero = tf.zeros([100,1])

ax.plot(x, mae_loss(x.reshape([100,1]),zero),
        label='ℓ¹-error')
ax.legend()

plt.savefig("../../plots/l1loss.pdf", bbox_inches='tight')
```

Source code 5.3 (code/loss_functions/l1loss_plot.py): PYTHON code used to create Figure 5.3

5.4.2 Mean squared error loss

Definition 5.4.2. *Let $d \in \mathbb{N}$ and let $\|\|\cdot\|\|\colon \mathbb{R}^d \to [0,\infty)$ be a norm. Then we say that $\mathbf{L}$ is the mean squared error loss function based on $\|\|\cdot\|\|$ if and only if it holds that $\mathbf{L}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 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*

$$\mathbf{L}(x,y) = \|\|x-y\|\|^2. \tag{5.40}$$

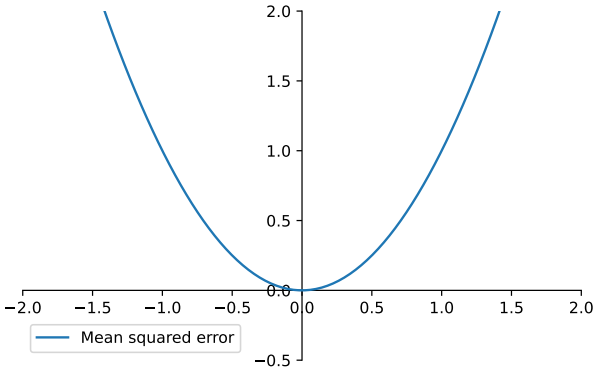

Figure 5.4 (plots/mseloss.pdf): A plot of the function $\mathbb{R} \ni x \mapsto \mathbf{L}(x,0) \in [0,\infty)$ where $\mathbf{L}$ is the mean squared error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0,\infty)$ (cf. Definition 5.4.2).

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-2,2), (-.5,2))

x = np.linspace(-2, 2, 100)

mse_loss = tf.keras.losses.MeanSquaredError(
    reduction=tf.keras.losses.Reduction.NONE)
zero = tf.zeros([100,1])

ax.plot(x, mse_loss(x.reshape([100,1]),zero),
        label='Mean squared error')
ax.legend()

plt.savefig("../../plots/mseloss.pdf", bbox_inches='tight')
```

Source code 5.4 (code/loss_functions/mseloss_plot.py): PYTHON code used to create Figure 5.4

Lemma 5.4.3. *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 5.4.2). 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

$$\mathbf{L}(u,v) = \mathbf{L}(x,y) + \mathbf{L}'(x,y)(u-x, v-y) + \tfrac{1}{2}\mathbf{L}^{(2)}(x,y)\big((u-x,v-y),(u-x,v-y)\big). \tag{5.41}$$

Proof of Lemma 5.4.3. Observe that (5.40) implies that for all $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$ it holds that

$$\mathbf{L}(x,y) = \|x-y\|_2^2 = \langle x-y, x-y\rangle = \sum_{i=1}^{d} (x_i - y_i)^2. \tag{5.42}$$

Therefore, we obtain that for all $x, y \in \mathbb{R}^d$ it holds that $\mathbf{L} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ and

$$(\nabla \mathbf{L})(x,y) = (2(x-y), -2(x-y)) \in \mathbb{R}^{2d}. \tag{5.43}$$

This implies that for all $x, y, h, k \in \mathbb{R}^d$ it holds that

$$\mathbf{L}'(x,y)(h,k) = \langle 2(x-y), h\rangle + \langle -2(x,y), k\rangle = 2\langle x-y, h-k\rangle. \tag{5.44}$$

Furthermore, note that (5.43) implies that for all $x, y \in \mathbb{R}^d$ it holds that $\mathbf{L} \in C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ and

$$(\mathrm{Hess}_{(x,y)}\, \mathbf{L}) = \begin{pmatrix} 2\,\mathrm{I}_d & -2\,\mathrm{I}_d \\ -2\,\mathrm{I}_d & 2\,\mathrm{I}_d \end{pmatrix}. \tag{5.45}$$

Hence, we obtain that for all $x, y, h, k \in \mathbb{R}^d$ it holds that

$$\mathbf{L}^{(2)}(x,y)\big((h,k),(h,k)\big) = 2\langle h,h\rangle - 2\langle h,k\rangle - 2\langle k,h\rangle + 2\langle k,k\rangle = 2\|h-k\|_2^2. \tag{5.46}$$

Combining this with (5.43) shows that for all $x, y \in \mathbb{R}^d$, $h, k \in \mathbb{R}^d$ it holds that $\mathbf{L} \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ and

$$\begin{aligned} &\mathbf{L}(x,y) + \mathbf{L}'(x,y)(h,k) + \tfrac{1}{2}\mathbf{L}^{(2)}(x,y)\big((h,k),(h,k)\big) \\ &= \|x-y\|_2^2 + 2\langle x-y, h-k\rangle + \|h-k\|_2^2 \\ &= \|x-y+(h-k)\|_2^2 \\ &= \mathbf{L}(x+h, y+k). \end{aligned} \tag{5.47}$$

This implies items (i) and (ii). The proof of Lemma 5.4.3 is thus complete. □

5.4.3 Huber error loss

Definition 5.4.4. *Let $d \in \mathbb{N}$, $\delta \in [0, \infty)$ and let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be a norm. Then we say that $\mathbf{L}$ is the δ-Huber-error loss function based on $\|\cdot\|$ if and only if it holds that $\mathbf{L} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 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*

$$\mathbf{L}(x, y) = \begin{cases} \frac{1}{2}\|x - y\|^2 & : \|x - y\| \leq \delta \\ \delta(\|x - y\| - \frac{\delta}{2}) & : \|x - y\| > \delta. \end{cases} \tag{5.48}$$

Lemma 5.4.5. *Let $\delta \in [0, \infty)$ and let $\mathbf{H} \colon \mathbb{R} \to [0, \infty)$ satisfy for all $z \in \mathbb{R}$ that*

$$\mathbf{H}(z) = \begin{cases} \frac{1}{2}z^2 & : z \leq \delta \\ \delta(z - \frac{\delta}{2}) & : z > \delta. \end{cases} \tag{5.49}$$

Then $\mathbf{H}$ is continuous.

Proof of Lemma 5.4.5. Throughout this proof, let $f, g \in C(\mathbb{R}, \mathbb{R})$ satisfy for all $z \in \mathbb{R}$ that

$$f(z) = \tfrac{1}{2}z^2 \qquad \text{and} \qquad g(z) = \delta(z - \tfrac{\delta}{2}). \tag{5.50}$$

Observe that (5.50) implies that

$$g(\delta) = \delta(\delta - \tfrac{\delta}{2}) = \tfrac{1}{2}\delta^2 = f(\delta). \tag{5.51}$$

Combining this with the fact that for all $z \in \mathbb{R}$ it holds that

$$\mathbf{H}(z) = \begin{cases} f(z) & : z \leq \delta \\ g(z) & : z > \delta \end{cases} \tag{5.52}$$

establishes that $\mathbf{H}$ is continuous. The proof of Lemma 5.4.5 is thus complete. □

Corollary 5.4.6. *Let $d \in \mathbb{N}$, $\delta \in [0, \infty)$, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be a norm, and let $\mathbf{L}$ be the δ-Huber-error loss function based on $\|\cdot\|$ (cf. Definition 5.4.4). Then $\mathbf{L}$ is continuous.*

Proof of Corollary 5.4.6. Throughout this proof, let $\mathbf{H} \colon \mathbb{R} \to [0, \infty)$ satisfy for all $z \in \mathbb{R}$ that

$$\mathbf{H}(z) = \begin{cases} \frac{1}{2}z^2 & : z \leq \delta \\ \delta(z - \frac{\delta}{2}) & : z > \delta. \end{cases} \tag{5.53}$$

Note that (5.48) demonstrates that for all $x, y \in \mathbb{R}^d$ it holds that

$$\mathbf{L}(x, y) = \mathbf{H}(\|x - y\|). \tag{5.54}$$

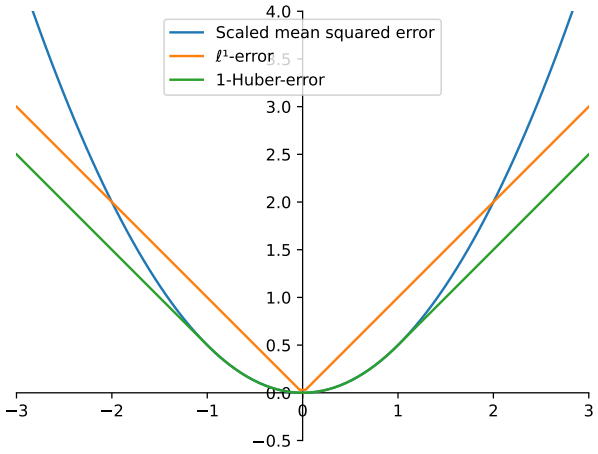

Figure 5.5 (plots/huberloss.pdf): A plot of the functions $\mathbb{R} \ni x \mapsto \mathbf{L}_i(x, 0) \in [0, \infty)$, $i \in \{1, 2, 3\}$, where $\mathbf{L}_0$ is the mean squared error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$, where $\mathbf{L}_1 \colon \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ satisfies for all $x, y \in \mathbb{R}^d$ that $\mathbf{L}_1(x, y) = \frac{1}{2}\mathbf{L}_0(x, y)$, where $\mathbf{L}_2$ is the ℓ^1-error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$, and where $\mathbf{L}_3$ is the 1-Huber loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$.

Furthermore, observe that Lemma 5.4.5 ensures that $\mathbf{H}$ is continuous. Combining this and the fact that $(\mathbb{R}^d \times \mathbb{R}^d \ni (x, y) \mapsto \|\|x - y\|\| \in \mathbb{R})$ is continuous with (5.54) proves that $\mathbf{L}$ is continuous. The proof of Corollary 5.4.6 is thus complete. □

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((-3,3), (-.5,4))

x = np.linspace(-3, 3, 100)

mse_loss = tf.keras.losses.MeanSquaredError(
    reduction=tf.keras.losses.Reduction.NONE)
mae_loss = tf.keras.losses.MeanAbsoluteError(
    reduction=tf.keras.losses.Reduction.NONE)
huber_loss = tf.keras.losses.Huber(
    reduction=tf.keras.losses.Reduction.NONE)

zero = tf.zeros([100,1])

ax.plot(x, mse_loss(x.reshape([100,1]),zero)/2.,
        label='Scaled mean squared error')
ax.plot(x, mae_loss(x.reshape([100,1]),zero),
```

```
22          label='ℓ¹-error')
23 ax.plot(x, huber_loss(x.reshape([100,1]),zero),
24          label='1-Huber-error')
25 ax.legend()
26
27 plt.savefig("../../plots/huberloss.pdf", bbox_inches='tight')
```

Source code 5.5 (code/loss_functions/huberloss_plot.py): PYTHON code used to create Figure 5.5

5.4.4 Cross-entropy loss

Definition 5.4.7. *Let $d \in \mathbb{N}$. Then we say that $\mathbf{L}$ is the d-dimensional cross-entropy loss function if and only if it holds that $\mathbf{L}\colon [0,\infty)^d \times [0,\infty)^d \to (-\infty,\infty]$ is the function from $[0,\infty)^d \times [0,\infty)^d$ to $(-\infty,\infty]$ which satisfies for all $x = (x_1,\dots,x_d)$, $y = (y_1,\dots,y_d) \in [0,\infty)^d$ that*

$$\mathbf{L}(x,y) = -\sum_{i=1}^{d} \lim_{\mathbf{x}\searrow x_i} \big[\ln(\mathbf{x}) y_i\big]. \tag{5.55}$$

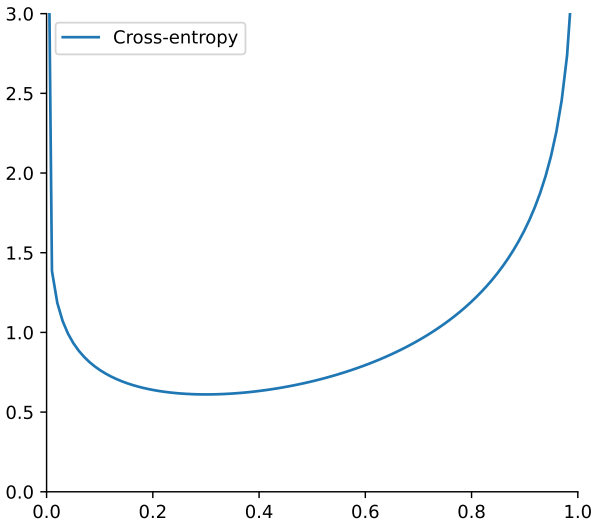

Figure 5.6 (plots/crossentropyloss.pdf): A plot of the function $(0,1) \ni x \mapsto \mathbf{L}\big((x,1-x),\big(\frac{3}{10},\frac{7}{10}\big)\big) \in \mathbb{R}$ where $\mathbf{L}$ is the 2-dimensional cross-entropy loss function (cf. Definition 5.4.7).

```
1 import numpy as np
2 import tensorflow as tf
```

```python
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((0,1), (0,3))

ax.set_aspect(.3)

x = np.linspace(0, 1, 100)

cce_loss = tf.keras.losses.CategoricalCrossentropy(
    reduction=tf.keras.losses.Reduction.NONE)
y = tf.constant([[0.3, 0.7]] * 100, shape=(100, 2))

X = tf.stack([x,1-x], axis=1)

ax.plot(x, cce_loss(y,X), label='Cross-entropy')
ax.legend()

plt.savefig("../../plots/crossentropyloss.pdf", bbox_inches='tight'
    )
```

Source code 5.6 (code/loss_functions/crossentropyloss_plot.py): Python code used to create Figure 5.6

Lemma 5.4.8. *Let $d \in \mathbb{N}$ and let $\mathbf{L}$ be the d-dimensional cross-entropy loss function (cf. Definition 5.4.7). Then*

(i) it holds for all $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in [0, \infty)^d$ that

$$(\mathbf{L}(x, y) = \infty) \leftrightarrow \big(\exists\, i \in \{1, 2, \ldots, d\} \colon [(x_i = 0) \wedge (y_i \neq 0)]\big), \tag{5.56}$$

(ii) it holds for all $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in [0, \infty)^d$ with $\forall\, i \in \{1, 2, \ldots, d\} \colon [(x_i \neq 0) \vee (y_i = 0)]$ that

$$\mathbf{L}(x, y) = -\sum_{\substack{i \in \{1,2,\ldots,d\}, \\ y_i \neq 0}} \ln(x_i) y_i \in \mathbb{R}, \tag{5.57}$$

and

(iii) it holds for all $x = (x_1, \ldots, x_d) \in (0, \infty)^d$, $y = (y_1, \ldots, y_d) \in [0, \infty)^d$ that

$$\mathbf{L}(x, y) = -\sum_{i=1}^{d} \ln(x_i) y_i \in \mathbb{R}. \tag{5.58}$$

Proof of Lemma 5.4.8. Note that (5.55) and the fact that for all $a, b \in [0, \infty)$ it holds that

$$\lim_{\mathbf{a} \searrow a} \big[\ln(\mathbf{a})b\big] = \begin{cases} 0 & : b = 0 \\ \ln(a)b & : (a \neq 0) \wedge (b \neq 0) \\ -\infty & : (a = 0) \wedge (b \neq 0) \end{cases} \tag{5.59}$$

establish items (i), (ii), and (iii). The proof of Lemma 5.4.8 is thus complete. □

Lemma 5.4.9. *Let $d \in \mathbb{N}$, let $\mathbf{L}$ be the d-dimensional cross-entropy loss function, let $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in [0, \infty)^d$ satisfy $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ and $x \neq y$, and let $f \colon [0, 1] \to (-\infty, \infty]$ satisfy for all $h \in [0, 1]$ that*

$$f(h) = \mathbf{L}(x + h(y - x), y) \tag{5.60}$$

(cf. Definition 5.4.7). Then f is strictly decreasing.

Proof of Lemma 5.4.9. Throughout this proof, let $g \colon [0, 1) \to (-\infty, \infty]$ satisfy for all $h \in [0, 1)$ that

$$g(h) = f(1 - h) \tag{5.61}$$

and let $J = \{i \in \{1, 2, \ldots, d\} \colon y_i \neq 0\}$. Observe that (5.60) shows that for all $h \in [0, 1)$ it holds that

$$g(h) = \mathbf{L}(x + (1 - h)(y - x), y) = \mathbf{L}(y + h(x - y), y). \tag{5.62}$$

Furthermore, note that the fact that for all $i \in J$ it holds that $x_i \in [0, \infty)$ and $y_i \in (0, \infty)$ implies that for all $i \in J$, $h \in [0, 1)$ it holds that

$$y_i + h(x_i - y_i) = (1 - h)y_i + hx_i \geq (1 - h)y_i > 0. \tag{5.63}$$

This, (5.62), and item (ii) in Lemma 5.4.8 demonstrate that for all $h \in [0, 1)$ it holds that

$$g(h) = -\sum_{i \in J} \ln(y_i + h(x_i - y_i))y_i \in \mathbb{R}. \tag{5.64}$$

The chain rule therefore ensures that for all $h \in [0, 1)$ it holds that $([0, 1) \ni z \mapsto g(z) \in \mathbb{R}) \in C^\infty([0, 1), \mathbb{R})$ and

$$g'(h) = -\sum_{i \in J} \frac{y_i(x_i - y_i)}{y_i + h(x_i - y_i)}. \tag{5.65}$$

This and the chain rule prove that for all $h \in [0, 1)$ it holds that

$$g''(h) = \sum_{i \in J} \frac{y_i(x_i - y_i)^2}{(y_i + h(x_i - y_i))^2}. \tag{5.66}$$

Moreover, observe that the fact that for all $z = (z_1, \ldots, z_d) \in [0,\infty)^d$ with $\sum_{i=1}^d z_i = \sum_{i=1}^d y_i$ and $\forall\, i \in J\colon z_i = y_i$ it holds that

$$\begin{aligned}\sum_{i\in\{1,2,\ldots,d\}\setminus J} z_i &= \left[\sum_{i\in\{1,2,\ldots,d\}} z_i\right] - \left[\sum_{i\in J} z_i\right] \\ &= \left[\sum_{i\in\{1,2,\ldots,d\}} y_i\right] - \left[\sum_{i\in J} z_i\right] \\ &= \sum_{i\in J}(y_i - z_i) = 0\end{aligned} \tag{5.67}$$

establishes that for all $z = (z_1, \ldots, z_d) \in [0,\infty)^d$ with $\sum_{i=1}^d z_i = \sum_{i=1}^d y_i$ and $\forall\, i \in J\colon z_i = y_i$ it holds that $z = y$. The assumption that $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ and $x \neq y$ hence implies that there exists $i \in J$ such that $x_i \neq y_i > 0$. Combining this with (5.66) shows that for all $h \in [0,1)$ it holds that

$$g''(h) > 0. \tag{5.68}$$

The fundamental theorem of calculus therefore demonstrates that for all $h \in (0,1)$ it holds that

$$g'(h) = g'(0) + \int_0^h g''(\mathfrak{h})\,\mathrm{d}\mathfrak{h} > g'(0). \tag{5.69}$$

In addition, note that (5.65) and the assumption that $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ ensure that

$$\begin{aligned}g'(0) &= -\sum_{i\in J}\frac{y_i(x_i - y_i)}{y_i} = \sum_{i\in J}(y_i - x_i) = \left[\sum_{i\in J} y_i\right] - \left[\sum_{i\in J} x_i\right] \\ &= \left[\sum_{i\in\{1,2,\ldots,d\}} y_i\right] - \left[\sum_{i\in J} x_i\right] = \left[\sum_{i\in\{1,2,\ldots,d\}} x_i\right] - \left[\sum_{i\in J} x_i\right] = \left[\sum_{i\in\{1,2,\ldots,d\}\setminus J} x_i\right] \geq 0.\end{aligned} \tag{5.70}$$

Combining this and (5.69) proves that for all $h \in (0,1)$ it holds that

$$g'(h) > 0. \tag{5.71}$$

Hence, we obtain that g is strictly increasing. This and (5.61) establish that $f|_{(0,1]}$ is strictly decreasing. Next observe that (5.61) and (5.64) imply that for all $h \in (0,1]$ it holds that

$$f(h) = -\sum_{i\in J}\ln(y_i + (1-h)(x_i - y_i))y_i = -\sum_{i\in J}\ln(x_i + h(y_i - x_i))y_i \in \mathbb{R}. \tag{5.72}$$

In the remainder of our proof that f is strictly decreasing we distinguish between the case $f(0) = \infty$ and the case $f(0) < \infty$. We first prove that f is strictly decreasing in the case

$$f(0) = \infty. \tag{5.73}$$

Note that (5.73), the fact that $f|_{(0,1]}$ is strictly decreasing, and (5.72) show that f is strictly decreasing. This establishes that f is strictly decreasing in the case $f(0) = \infty$. In the next step we prove that f is strictly decreasing in the case

$$f(0) < \infty. \tag{5.74}$$

Observe that (5.74) and items (i) and (ii) in Lemma 5.4.8 demonstrate that

$$0 \notin \cup_{i\in J}\{x_i\} \qquad \text{and} \qquad f(0) = -\sum_{i\in J} \ln(x_i + 0(y_i - x_i))y_i \in \mathbb{R}. \tag{5.75}$$

Combining this with (5.72) ensures that $f([0,1]) \subseteq \mathbb{R}$ and

$$([0,1] \ni h \mapsto f(h) \in \mathbb{R}) \in C([0,1], \mathbb{R}). \tag{5.76}$$

This and the fact that $f|_{(0,1]}$ is strictly decreasing prove that f is strictly decreasing. This establishes that f is strictly decreasing in the case $f(0) < \infty$. The proof of Lemma 5.4.9 is thus complete. □

Corollary 5.4.10. *Let $d \in \mathbb{N}$, let $A = \{x = (x_1, \ldots, x_d) \in [0,1]^d \colon \sum_{i=1}^d x_i = 1\}$, let $\mathbf{L}$ be the d-dimensional cross-entropy loss function, and let $y \in A$ (cf. Definition 5.4.7). Then*

(i) it holds that

$$\big\{x \in A \colon \mathbf{L}(x,y) = \inf\nolimits_{z\in A} \mathbf{L}(z,y)\big\} = \{y\} \tag{5.77}$$

and

(ii) it holds that

$$\inf_{z\in A} \mathbf{L}(z,y) = \mathbf{L}(y,y) = -\sum_{\substack{i\in\{1,2,\ldots,d\},\\ y_i\neq 0}} \ln(y_i)y_i. \tag{5.78}$$

Proof of Corollary 5.4.10. Note that Lemma 5.4.9 shows that for all $x \in A\backslash\{y\}$ it holds that

$$\mathbf{L}(x,y) = \mathbf{L}(x + 0(y-x), y) > \mathbf{L}(x + 1(y-x), y) = \mathbf{L}(y,y). \tag{5.79}$$

This and item (ii) in Lemma 5.4.8 establish items (i) and (ii). The proof of Corollary 5.4.10 is thus complete. □

5.4.5 Kullback–Leibler divergence loss

Lemma 5.4.11. *Let $z \in (0, \infty)$. Then*

(i) it holds that

$$\liminf_{x \searrow 0} |\ln(x)x| = 0 \tag{5.80}$$

and

(ii) it holds for all $y \in [0, \infty)$ that

$$\liminf_{\mathbf{y} \searrow y} \left[\ln\left(\tfrac{z}{\mathbf{y}}\right)\mathbf{y}\right] = \limsup_{\mathbf{y} \searrow y} \left[\ln\left(\tfrac{z}{\mathbf{y}}\right)\mathbf{y}\right] = \begin{cases} 0 & : y = 0 \\ \ln\left(\frac{z}{y}\right) y & : y > 0 \end{cases} \tag{5.81}$$

Proof of Lemma 5.4.11. Throughout this proof, let $f \colon (0, \infty) \to \mathbb{R}$ and $g \colon (0, \infty) \to \mathbb{R}$ satisfy for all $x \in (0, \infty)$ that

$$f(x) = \ln(x^{-1}) \qquad \text{and} \qquad g(x) = x. \tag{5.82}$$

Observe that the chain rule implies that for all $x \in (0, \infty)$ it holds that f is differentiable and

$$f'(x) = -x^{-2}(x^{-1})^{-1} = -x^{-1}. \tag{5.83}$$

Combining this, the fact that $\lim_{x \to \infty} |f(x)| = \infty = \lim_{x \to \infty} |g(x)|$, the fact that g is differentiable, the fact that for all $x \in (0, \infty)$ it holds that $g'(x) = 1 \neq 0$, and the fact that $\lim_{x \to \infty} \frac{-x^{-1}}{1} = 0$ with l'Hôpital's rule shows that

$$\liminf_{x \to \infty} \tfrac{f(x)}{g(x)} = 0 = \limsup_{x \to \infty} \tfrac{f(x)}{g(x)}. \tag{5.84}$$

This demonstrates that

$$\liminf_{x \searrow 0} \tfrac{f(x^{-1})}{g(x^{-1})} = 0 = \limsup_{x \searrow 0} \tfrac{f(x^{-1})}{g(x^{-1})}. \tag{5.85}$$

The fact that for all $x \in (0, \infty)$ it holds that $\frac{f(x^{-1})}{g(x^{-1})} = \ln(x)x$ therefore proves item (i). Note that item (i) and the fact that for all $x \in (0, \infty)$ it holds that $\ln\left(\frac{z}{x}\right)x = \ln(z)x - \ln(x)x$ establish item (ii). The proof of Lemma 5.4.11 is thus complete. □

Definition 5.4.12. *Let $d \in \mathbb{N}$. Then we say that $\mathbf{L}$ is the d-dimensional Kullback–Leibler divergence loss function if and only if it holds that $\mathbf{L} \colon [0, \infty)^d \times [0, \infty)^d \to (-\infty, \infty]$ is the function from $[0, \infty)^d \times [0, \infty)^d$ to $(-\infty, \infty]$ which satisfies for all $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in [0, \infty)^d$ that*

$$\mathbf{L}(x, y) = -\sum_{i=1}^{d} \lim_{\mathbf{x} \searrow x_i} \lim_{\mathbf{y} \searrow y_i} \left[\ln\left(\tfrac{\mathbf{x}}{\mathbf{y}}\right)\mathbf{y}\right] \tag{5.86}$$

(cf. Lemma 5.4.11).

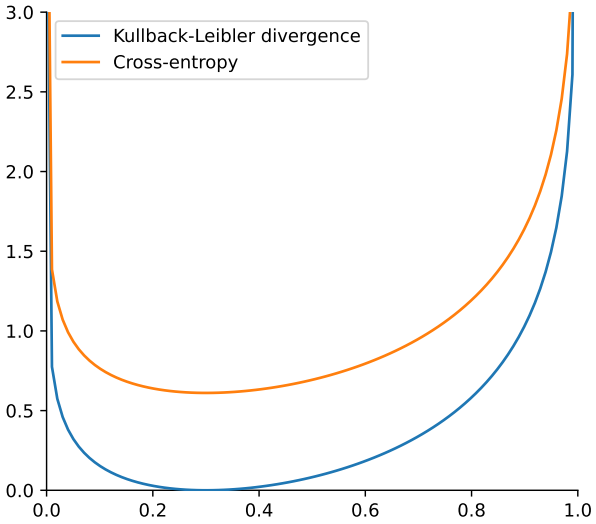

Figure 5.7 (plots/kldloss.pdf): A plot of the functions $(0,1) \ni x \mapsto \mathbf{L}_i\big((x, 1-x), \big(\frac{3}{10}, \frac{7}{10}\big)\big) \in \mathbb{R}$, $i \in \{1,2\}$, where $\mathbf{L}_1$ is the 2-dimensional Kullback–Leibler divergence loss function and where $\mathbf{L}_2$ is the 2-dimensional cross-entropy loss function (cf. Definitions 5.4.7 and 5.4.12).

```
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import plot_util

ax = plot_util.setup_axis((0,1), (0,3))

ax.set_aspect(.3)

x = np.linspace(0, 1, 100)

kld_loss = tf.keras.losses.KLDivergence(
    reduction=tf.keras.losses.Reduction.NONE)
cce_loss = tf.keras.losses.CategoricalCrossentropy(
    reduction=tf.keras.losses.Reduction.NONE)
y = tf.constant([[0.3, 0.7]] * 100, shape=(100, 2))

X = tf.stack([x,1-x], axis=1)

ax.plot(x, kld_loss(y,X), label='Kullback-Leibler divergence')
ax.plot(x, cce_loss(y,X), label='Cross-entropy')
ax.legend()
```

```
23
24 plt.savefig("../../plots/kldloss.pdf", bbox_inches='tight')
```

Source code 5.7 (code/loss_functions/kldloss_plot.py): PYTHON code used to create Figure 5.7

Lemma 5.4.13. *Let $d \in \mathbb{N}$, let $\mathbf{L}_{\mathrm{CE}}$ be the d-dimensional cross-entropy loss function, and let $\mathbf{L}_{\mathrm{KLD}}$ be the d-dimensional Kullback–Leibler divergence loss function (cf. Definitions 5.4.7 and 5.4.12). Then it holds for all $x, y \in [0,\infty)^d$ that*

$$\mathbf{L}_{\mathrm{CE}}(x,y) = \mathbf{L}_{\mathrm{KLD}}(x,y) + \mathbf{L}_{\mathrm{CE}}(y,y). \tag{5.87}$$

Proof of Lemma 5.4.13. Observe that Lemma 5.4.11 ensures that for all $a, b \in [0,\infty)$ it holds that

$$\begin{aligned}\lim_{\mathbf{a}\searrow a}\lim_{\mathbf{b}\searrow b}\big[\ln\big(\tfrac{\mathbf{a}}{\mathbf{b}}\big)\mathbf{b}\big] &= \lim_{\mathbf{a}\searrow a}\lim_{\mathbf{b}\searrow b}\big[\ln(\mathbf{a})\mathbf{b} - \ln(\mathbf{b})\mathbf{b}\big] \\ &= \lim_{\mathbf{a}\searrow a}\Big[\ln(\mathbf{a})b - \lim_{\mathbf{b}\searrow b}[\ln(\mathbf{b})\mathbf{b}]\Big] \\ &= \Big(\lim_{\mathbf{a}\searrow a}[\ln(\mathbf{a})b]\Big) - \Big(\lim_{\mathbf{b}\searrow b}[\ln(\mathbf{b})\mathbf{b}]\Big).\end{aligned} \tag{5.88}$$

This and (5.86) imply that for all $x = (x_1,\dots,x_d)$, $y = (y_1,\dots,y_d) \in [0,\infty)^d$ it holds that

$$\begin{aligned}\mathbf{L}_{\mathrm{KLD}}(x,y) &= -\sum_{i=1}^{d}\lim_{\mathbf{x}\searrow x_i}\lim_{\mathbf{y}\searrow y_i}\big[\ln\big(\tfrac{\mathbf{x}}{\mathbf{y}}\big)\mathbf{y}\big] \\ &= -\left(\sum_{i=1}^{d}\lim_{\mathbf{x}\searrow x_i}[\ln(\mathbf{x})y_i]\right) + \left(\sum_{i=1}^{d}\lim_{\mathbf{y}\searrow y_i}[\ln(\mathbf{y})\mathbf{y}]\right).\end{aligned} \tag{5.89}$$

Furthermore, note that Lemma 5.4.11 shows that for all $b \in [0,\infty)$ it holds that

$$\lim_{\mathbf{b}\searrow b}\big[\ln(\mathbf{b})\mathbf{b}\big] = \begin{cases} 0 & : b = 0 \\ \ln(b)b & : b > 0 \end{cases} = \lim_{\mathbf{b}\searrow b}\big[\ln(\mathbf{b})b\big]. \tag{5.90}$$

Combining this with (5.89) demonstrates that for all $x = (x_1,\dots,x_d)$, $y = (y_1,\dots,y_d) \in [0,\infty)^d$ it holds that

$$\mathbf{L}_{\mathrm{KLD}}(x,y) = -\left(\sum_{i=1}^{d}\lim_{\mathbf{x}\searrow x_i}[\ln(\mathbf{x})y_i]\right) + \left(\sum_{i=1}^{d}\lim_{\mathbf{y}\searrow y_i}[\ln(\mathbf{y})\mathbf{y}]\right) = \mathbf{L}_{\mathrm{CE}}(x,y) - \mathbf{L}_{\mathrm{CE}}(y,y). \tag{5.91}$$

Hence, we obtain (5.87). The proof of Lemma 5.4.13 is thus complete. □

Lemma 5.4.14. *Let $d \in \mathbb{N}$, let $\mathbf{L}$ be the d-dimensional Kullback–Leibler divergence loss function, let $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in [0, \infty)^d$ satisfy $\sum_{i=1}^{d} x_i = \sum_{i=1}^{d} y_i$ and $x \neq y$, and let $f \colon [0,1] \to (-\infty, \infty]$ satisfy for all $h \in [0,1]$ that*

$$f(h) = \mathbf{L}(x + h(y - x), y) \tag{5.92}$$

(cf. Definition 5.4.12). Then f is strictly decreasing.

Proof of Lemma 5.4.14. Observe that Lemma 5.4.9 and Lemma 5.4.13 prove that f is strictly decreasing. The proof of Lemma 5.4.14 is thus complete. □

Corollary 5.4.15. *Let $d \in \mathbb{N}$, let $A = \{x = (x_1, \ldots, x_d) \in [0,1]^d \colon \sum_{i=1}^{d} x_i = 1\}$, let $\mathbf{L}$ be the d-dimensional Kullback–Leibler divergence loss function, and let $y \in A$ (cf. Definition 5.4.12). Then*

(i) it holds that

$$\big\{x \in A \colon \mathbf{L}(x, y) = \inf\nolimits_{z \in A} \mathbf{L}(z, y)\big\} = \{y\} \tag{5.93}$$

and

(ii) it holds that $\inf_{z \in A} \mathbf{L}(z, y) = \mathbf{L}(y, y) = 0$.

Proof of Corollary 5.4.15. Note that Lemma 5.4.13 and Lemma 5.4.13 establish items (i) and (ii). The proof of Corollary 5.4.15 is thus complete. □

5.5 GF optimization in the training of ANNs

Example 5.5.1. *Let $d, L, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \ldots, l_L \in \mathbb{N}$ satisfy*

$$\mathfrak{d} = l_1(d + 1) + \Big[\sum_{k=2}^{L} l_k(l_{k-1} + 1)\Big], \tag{5.94}$$

let $a \colon \mathbb{R} \to \mathbb{R}$ be continuously differentiable, let $M \in \mathbb{N}$, $x_1, x_2, \ldots, x_M \in \mathbb{R}^d$, $y_1, y_2, \ldots, y_M \in \mathbb{R}^{l_L}$, let $\mathbf{L} \colon \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \to \mathbb{R}$ be the mean squared error loss function based on $\mathbb{R}^d \ni x \mapsto \|x\|_2 \in [0, \infty)$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbf{L}\big(\big(\mathcal{N}^{\theta,d}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \ldots, \mathfrak{M}_{a,l_{L-1}}, \mathrm{id}_{\mathbb{R}^{l_L}}}\big)(x_m), y_m\big)\right], \tag{5.95}$$

let $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, \infty)$ that

$$\Theta_t = \xi - \int_0^t (\nabla \mathscr{L})(\Theta_s)\, \mathrm{d}s \tag{5.96}$$

(cf. Definitions 1.1.3, 1.2.1, 3.3.4, and 5.4.2, Corollary 5.3.6, and Lemma 5.4.3). Then Θ *is a GF trajectory for the objective function* $\mathscr{L}$ *with initial value* ξ *(cf. Definition 5.2.1).*

Proof for Example 5.5.1. Observe that (5.9), (5.10), and (5.96) demonstrate that Θ is a GF trajectory for the objective function $\mathscr{L}$ with initial value ξ (cf. Definition 5.2.1). The proof for Example 5.5.1 is thus complete. □

Example 5.5.2. *Let* $d, L, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \ldots, l_L \in \mathbb{N}$ *satisfy*

$$\mathfrak{d} = l_1(d+1) + \Big[\sum_{k=2}^{L} l_k(l_{k-1}+1)\Big], \tag{5.97}$$

let $a \colon \mathbb{R} \to \mathbb{R}$ *be continuously differentiable, let* $A \colon \mathbb{R}^{l_L} \to \mathbb{R}^{l_L}$ *be the* l_L*-dimensional softmax activation function, let* $M \in \mathbb{N}$, $x_1, x_2, \ldots, x_M \in \mathbb{R}^d$, $y_1, y_2, \ldots, y_M \in [0,\infty)^{l_L}$, *let* $\mathbf{L}_1$ *be the* l_L*-dimensional cross-entropy loss function, let* $\mathbf{L}_2$ *be the* l_L*-dimensional Kullback–Leibler divergence loss function, for every* $i \in \{1,2\}$ *let* $\mathscr{L}_i \colon \mathbb{R}^{\mathfrak{d}} \to [0,\infty)$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$\mathscr{L}_i(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \mathbf{L}_i\big(\big(\mathcal{N}^{\theta,d}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_{L-1}},A}\big)(x_m), y_m\big)\right], \tag{5.98}$$

let $\xi \in \mathbb{R}^{\mathfrak{d}}$, *and for every* $i \in \{1,2\}$ *let* $\Theta^i \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$ *satisfy for all* $t \in [0,\infty)$ *that*

$$\Theta^i_t = \xi - \int_0^t (\nabla \mathscr{L}_i)(\Theta^i_s)\,\mathrm{d}s \tag{5.99}$$

(cf. Definitions 1.1.3, 1.2.1, 1.2.43, 5.4.7, and 5.4.12 and Corollary 5.3.7). Then it holds for all $i, j \in \{1,2\}$ *that* Θ^i *is a GF trajectory for the objective function* $\mathscr{L}_j$ *with initial value* ξ *(cf. Definition 5.2.1).*

Proof for Example 5.5.2. Note that Lemma 5.4.13 ensures that for all $x, y \in (0,\infty)^{l_L}$ it holds that

$$(\nabla_x \mathbf{L}_1)(x,y) = (\nabla_x \mathbf{L}_2)(x,y). \tag{5.100}$$

Therefore, we obtain that for all $x \in \mathbb{R}^d$ it holds that

$$(\nabla \mathscr{L}_1)(x) = (\nabla \mathscr{L}_2)(x). \tag{5.101}$$

This, (5.9), (5.10), and (5.99) imply that for all $i \in \{1,2\}$ it holds that Θ^i is a GF trajectory for the objective function $\mathscr{L}_j$ with initial value ξ (cf. Definition 5.2.1). The proof for Example 5.5.2 is thus complete. □

5.6 Critical points in optimization problems

5.6.1 Local and global minimizers

Definition 5.6.1 (Local minimum point)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $\vartheta \in O$, and let $\mathscr{L} \colon O \to \mathbb{R}$ be a function. Then we say that ϑ is a local minimum point of $\mathscr{L}$ (we say that ϑ is a local minimizer of $\mathscr{L}$) if and only if there exists $\varepsilon \in (0, \infty)$ such that for all $\theta \in O$ with $\|\theta - \vartheta\|_2 < \varepsilon$ it holds that*

$$\mathscr{L}(\vartheta) \leq \mathscr{L}(\theta) \tag{5.102}$$

(cf. Definition 3.3.4).

Definition 5.6.2 (Global minimum point)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $\vartheta \in O$, and let $\mathscr{L} \colon O \to \mathbb{R}$ be a function. Then we say that ϑ is a global minimum point of $\mathscr{L}$ (we say that ϑ is a global minimizer of $\mathscr{L}$) if and only if it holds for all $\theta \in O$ that*

$$\mathscr{L}(\vartheta) \leq \mathscr{L}(\theta). \tag{5.103}$$

.

5.6.2 Local and global maximizers

Definition 5.6.3 (Local maximum point)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $\vartheta \in O$, and let $\mathscr{L} \colon O \to \mathbb{R}$ be a function. Then we say that ϑ is a local maximum point of $\mathscr{L}$ (we say that ϑ is a local maximizer of $\mathscr{L}$) if and only if there exists $\varepsilon \in (0, \infty)$ such that for all $\theta \in O$ with $\|\theta - \vartheta\|_2 < \varepsilon$ it holds that*

$$\mathscr{L}(\vartheta) \geq \mathscr{L}(\theta) \tag{5.104}$$

(cf. Definition 3.3.4).

Definition 5.6.4 (Global maximum point)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $\vartheta \in O$, and let $\mathscr{L} \colon O \to \mathbb{R}$ be a function. Then we say that ϑ is a global maximum point of $\mathscr{L}$ (we say that ϑ is a global maximizer of $\mathscr{L}$) if and only if it holds for all $\theta \in O$ that*

$$\mathscr{L}(\vartheta) \geq \mathscr{L}(\theta). \tag{5.105}$$

.

5.6.3 Critical points

Definition 5.6.5 (Critical point). *Let $\mathfrak{d} \in \mathbb{N}$, let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be an environment of ϑ, and let $\mathscr{L} \colon O \to \mathbb{R}$ be differentiable at ϑ. Then we say that ϑ is a critical point of $\mathscr{L}$ if and only if it holds that*

$$(\nabla \mathscr{L})(\vartheta) = 0. \tag{5.106}$$

Lemma 5.6.6. *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\vartheta \in O$, let $\mathscr{L} \colon O \to \mathbb{R}$ be a function, assume that $\mathscr{L}$ is differentiable at ϑ, and assume that $(\nabla \mathscr{L})(\vartheta) \neq 0$. Then there exists $\theta \in O$ such that $\mathscr{L}(\theta) < \mathscr{L}(\vartheta)$.*

Proof of Lemma 5.6.6. Throughout this proof, let $v \in \mathbb{R}^{\mathfrak{d}} \backslash \{0\}$ satisfy $v = -(\nabla \mathscr{L})(\vartheta)$, let $\delta \in (0, \infty)$ satisfy for all $t \in (-\delta, \delta)$ that

$$\vartheta + tv = \vartheta - t(\nabla \mathscr{L})(\vartheta) \in O, \tag{5.107}$$

and let $L \colon (-\delta, \delta) \to \mathbb{R}$ satisfy for all $t \in (-\delta, \delta)$ that

$$L(t) = \mathscr{L}(\vartheta + tv). \tag{5.108}$$

Note that for all $t \in (0, \delta)$ it holds that

$$\begin{aligned} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| &= \left| \left[\frac{\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta)}{t} \right] + \|(\nabla \mathscr{L})(\vartheta)\|_2^2 \right| \\ &= \left| \left[\frac{\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta)}{t} \right] + \langle (\nabla \mathscr{L})(\vartheta), (\nabla \mathscr{L})(\vartheta) \rangle \right| \\ &= \left| \left[\frac{\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta)}{t} \right] - \langle (\nabla \mathscr{L})(\vartheta), v \rangle \right|. \end{aligned} \tag{5.109}$$

Therefore, we obtain that for all $t \in (0, \delta)$ it holds that

$$\begin{aligned} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| &= \left| \left[\frac{\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta)}{t} \right] - \mathscr{L}'(\vartheta) v \right| \\ &= \left| \frac{\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta) - \mathscr{L}'(\vartheta) tv}{t} \right| = \frac{|\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta) - \mathscr{L}'(\vartheta) tv|}{t}. \end{aligned} \tag{5.110}$$

The assumption that $\mathscr{L}$ is differentiable at ϑ hence demonstrates that

$$\limsup_{t \searrow 0} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| = 0. \tag{5.111}$$

The fact that $\|v\|_2^2 > 0$ therefore demonstrates that there exists $t \in (0, \delta)$ which satisfies

$$\left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| < \frac{\|v\|_2^2}{2}. \tag{5.112}$$

Observe that the triangle inequality, the fact that $\|v\|_2^2 > 0$, and (5.112) prove that

$$\begin{aligned} \frac{L(t) - L(0)}{t} &= \left[\frac{L(t) - L(0)}{t} + \|v\|_2^2\right] - \|v\|_2^2 \leq \left|\left[\frac{L(t) - L(0)}{t}\right] + \|v\|_2^2\right| - \|v\|_2^2 \\ &< \frac{\|v\|_2^2}{2} - \|v\|_2^2 = -\frac{\|v\|_2^2}{2} < 0. \end{aligned} \tag{5.113}$$

This ensures that

$$\mathscr{L}(\vartheta + tv) = L(t) < L(0) = \mathscr{L}(\vartheta). \tag{5.114}$$

The proof of Lemma 5.6.6 is thus complete. □

Lemma 5.6.7 (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 $\mathscr{L} \colon O \to \mathbb{R}$ be a function, assume that $\mathscr{L}$ is differentiable at ϑ, and assume*

$$\mathscr{L}(\vartheta) = \inf_{\theta \in O} \mathscr{L}(\theta). \tag{5.115}$$

Then $(\nabla \mathscr{L})(\vartheta) = 0$.

Proof of Lemma 5.6.7. We prove Lemma 5.6.7 by contradiction. We thus assume that $(\nabla \mathscr{L})(\vartheta) \neq 0$. Lemma 5.6.6 then implies that there exists $\theta \in O$ such that $\mathscr{L}(\theta) < \mathscr{L}(\vartheta)$. Combining this with (5.115) shows that

$$\mathscr{L}(\theta) < \mathscr{L}(\vartheta) = \inf_{w \in O} \mathscr{L}(w) \leq \mathscr{L}(\theta). \tag{5.116}$$

The proof of Lemma 5.6.7 is thus complete. □

Corollary 5.6.8 (Necessary condition for local minimum points)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\vartheta \in O$, let $\mathscr{L} \colon O \to \mathbb{R}$ be differentiable at ϑ, and assume that ϑ is a local minimum point of $\mathscr{L}$. Then ϑ is a critical point of $\mathscr{L}$ (cf. Definition 5.6.5).*

Proof of Corollary 5.6.8. Note that Lemma 5.6.7 shows that $(\nabla \mathscr{L})(\vartheta) = 0$. The proof of Corollary 5.6.8 is thus complete. □

5.7 Conditions on objective functions in optimization problems

In this section we discuss different common assumptions from the scientific literature on the objective function (the function one intends to minimize) of optimization problems. For further reading we refer, for instance, to [148].

5.7.1 Convexity

Definition 5.7.1 (Convex functions). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function. Then we say that $\mathscr{L}$ is a convex function (we say that $\mathscr{L}$ is convex) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that*

$$\mathscr{L}(tv + (1-t)w) \leq t\mathscr{L}(v) + (1-t)\mathscr{L}(w). \tag{5.117}$$

Lemma 5.7.2 (Equivalence for convex functions). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$. Then the following three statements are equivalent:*

(i) It holds that $\mathscr{L}$ is convex (cf. Definition 5.7.1).

(ii) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$\mathscr{L}(\theta + tv) \leq \mathscr{L}(\theta) + t(\mathscr{L}(\theta + v) - \mathscr{L}(\theta)). \tag{5.118}$$

(iii) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$t\big(\mathscr{L}(\theta + v) - \mathscr{L}(\theta + tv)\big) - (1-t)\big(\mathscr{L}(\theta + tv) - \mathscr{L}(\theta)\big) \geq 0. \tag{5.119}$$

Proof of Lemma 5.7.2. Observe that (5.117) establishes that ((i) $\leftrightarrow$ (ii)) and ((i) $\leftrightarrow$ (iii)). The proof of Lemma 5.7.2 is thus complete. □

Lemma 5.7.3 (Equivalence for differentiable convex functions). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable. Then the following three statements are equivalent:*

(i) It holds that $\mathscr{L}$ is convex (cf. Definition 5.7.1).

(ii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle \tag{5.120}$$

(cf. Definition 1.4.7).

(iii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \geq 0 \tag{5.121}$$

(cf. Definition 1.4.7).

Proof of Lemma 5.7.3. We first prove that ((i) → (ii)). For this assume that $\mathscr{L}$ is convex (cf. Definition 5.7.1). C8.1Note that the assumption that $\mathscr{L}$ is convex and item (ii) in Lemma 5.7.2 demonstrate that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ it holds that

$$\mathscr{L}(w + t(v - w)) \leq \mathscr{L}(w) + t(\mathscr{L}(v) - \mathscr{L}(w)). \tag{5.122}$$

C5.1Hence, we obtain that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ it holds that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \frac{\mathscr{L}(w + t(v - w)) - \mathscr{L}(w)}{t}. \tag{5.123}$$

C3.2Combining this and the assumption that $\mathscr{L}$ is differentiable proves that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \limsup_{t \to 0} \frac{\mathscr{L}(w + t(v - w)) - \mathscr{L}(w)}{t} = \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle \tag{5.124}$$

(cf. Definition 1.4.7). This proves that ((i) → (ii)).

In the next step we prove that ((ii) → (iii)). For this assume that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle. \tag{5.125}$$

C8.1Observe that (5.125) establishes that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \mathscr{L}(v) + \mathscr{L}(w) &\geq \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle + \mathscr{L}(v) + \langle (\nabla \mathscr{L})(v), w - v \rangle \\ &= \mathscr{L}(v) + \mathscr{L}(w) - \langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \end{aligned} \tag{5.126}$$

C5.2This ensures that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \geq 0. \tag{5.127}$$

This proves that ((ii) → (iii)).

In the next step we prove that ((iii) → (i)). For this assume that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \geq 0. \tag{5.128}$$

C8.1Note that (5.128) implies that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\alpha, \beta \in \mathbb{R}$ with $\alpha > \beta$ it holds that

$$\begin{aligned} &\langle (\nabla \mathscr{L})(\theta + \alpha v) - (\nabla \mathscr{L})(\theta + \beta v), v \rangle \\ &= (\alpha - \beta)^{-1} \langle (\nabla \mathscr{L})(\theta + \alpha v) - (\nabla \mathscr{L})(\theta + \beta v), (\alpha - \beta) v \rangle \geq 0. \end{aligned} \tag{5.129}$$

C3.2Combining this and the fundamental theorem of calculus shows that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ it holds that

$$\begin{aligned}
&t\big(\mathscr{L}(\theta+v)-\mathscr{L}(\theta+tv)\big)-(1-t)\big(\mathscr{L}(\theta+tv)-\mathscr{L}(\theta)\big)\\
&= t\left(\int_t^1 \langle(\nabla\mathscr{L})(\theta+sv), v\rangle\, ds\right) - (1-t)\left(\int_0^t \langle(\nabla\mathscr{L})(\theta+sv), v\rangle\, ds\right)\\
&= t(1-t)\left(\int_0^1 \langle(\nabla\mathscr{L})(\theta+(t+s(1-t))v), v\rangle\, ds\right)\\
&\quad - (1-t)t\left(\int_0^1 \langle(\nabla\mathscr{L})(\theta+stv), v\rangle\, ds\right)\\
&= t(1-t)\left(\int_0^1 \langle(\nabla\mathscr{L})(\theta+(t+s(1-t))v) - (\nabla\mathscr{L})(\theta+stv), v\rangle\, ds\right)\\
&\geq 0.
\end{aligned} \tag{5.130}$$

C3.1This and item (iii) in Lemma 5.7.2 demonstrate that $\mathscr{L}$ is convex. This proves that ((iii) $\to$ (i)). The proof of Lemma 5.7.3 is thus complete. □

5.7.2 Monotonicity

Definition 5.7.4 (Monotonically increasing functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that $\mathscr{G}$ is a monotonically increasing function (we say that $\mathscr{G}$ is monotonically increasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \mathscr{G}(v) - \mathscr{G}(w), v - w\rangle \geq 0 \tag{5.131}$$

(cf. Definition 1.4.7).

Definition 5.7.5 (Monotonically decreasing functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that $\mathscr{G}$ is a monotonically decreasing function (we say that $\mathscr{G}$ is monotonically decreasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \mathscr{G}(v) - \mathscr{G}(w), v - w\rangle \leq 0 \tag{5.132}$$

(cf. Definition 1.4.7).

Lemma 5.7.6 (Equivalence for monotonically increasing and decreasing functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then the following two statements are equivalent:*

(i) It holds that $\mathscr{G}$ is monotonically increasing (cf. Definition 5.7.4).

(ii) It holds that $-\mathscr{G}$ is monotonically decreasing (cf. Definition 5.7.5).

Proof of Lemma 5.7.6. Observe that (5.131) and (5.132) prove that ((i) $\leftrightarrow$ (ii)). The proof of Lemma 5.7.6 is thus complete. □

Lemma 5.7.7 (Convexity and monotonicity)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable. Then the following three statements are equivalent:*

(i) It holds that $\mathscr{L}$ is convex (cf. Definition 5.7.1).

(ii) It holds that $\nabla \mathscr{L}$ is monotonically increasing (cf. Definition 5.7.4).

(iii) It holds that $-(\nabla \mathscr{L})$ is monotonically decreasing (cf. Definition 5.7.5).

Proof of Lemma 5.7.7. C8.1Note that Lemma 5.7.3 and Lemma 5.7.6 establish that ((i) $\leftrightarrow$ (ii)) and that ((i) $\leftrightarrow$ (iii)). The proof of Lemma 5.7.7 is thus complete. □

Definition 5.7.8 (Generalized monotonically increasing functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that $\mathscr{G}$ is a c-generalized monotonically increasing function (we say that $\mathscr{G}$ is c-generalized monotonically increasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \mathscr{G}(v) - \mathscr{G}(w), v - w \rangle \geq c \|v - w\|_2^2 \tag{5.133}$$

(cf. Definitions 1.4.7 and 3.3.4).

Definition 5.7.9 (Generalized monotonically decreasing functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that $\mathscr{G}$ is a c-generalized monotonically decreasing function (we say that $\mathscr{G}$ is c-generalized monotonically decreasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \mathscr{G}(v) - \mathscr{G}(w), v - w \rangle \leq -c \|v - w\|_2^2. \tag{5.134}$$

(cf. Definitions 1.4.7 and 3.3.4).

Lemma 5.7.10 (Equivalence for monotonically increasing and decreasing functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then the following two statements are equivalent:*

(i) *It holds that $\mathscr{G}$ is c-generalized monotonically increasing (cf. Definition 5.7.8).*

(ii) *It holds that $-\mathscr{G}$ is (-c)-generalized monotonically decreasing (cf. Definition 5.7.9).*

Proof of Lemma 5.7.10. Observe that (5.133) and (5.134) ensure that ((i) $\leftrightarrow$ (ii)). The proof of Lemma 5.7.10 is thus complete. □

5.7.3 Subgradients

Definition 5.7.11 (Subgradients)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $g, \theta \in \mathbb{R}^{\mathfrak{d}}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function. Then we say that g is a subgradient of $\mathscr{L}$ at θ if and only if it holds for all $v \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(v) \geq \mathscr{L}(\theta) + \langle g, v - \theta \rangle \tag{5.135}$$

(cf. Definition 1.4.7).

Lemma 5.7.12 (Convexity and subgradients)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable. Then the following two statements are equivalent:*

(i) *It holds that $\mathscr{L}$ is convex (cf. Definition 5.7.1).*

(ii) *It holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $(\nabla \mathscr{L})(\theta)$ is a subgradient of $\mathscr{L}$ at θ (cf. Definition 5.7.11).*

Proof of Lemma 5.7.12. C8.1Note that Lemma 5.7.3 proves that ((i) $\leftrightarrow$ (ii)). The proof of Lemma 5.7.12 is thus complete. □

5.7.4 Strong convexity

Definition 5.7.13 (Generalized convex functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function. Then we say that $\mathscr{L}$ is a c-generalized convex function (we say that $\mathscr{L}$ is c-generalized convex) if and only if it holds that*

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathscr{L}(\theta) - \tfrac{c}{2}\|\theta\|_2^2 \in \mathbb{R} \tag{5.136}$$

is convex (cf. Definitions 3.3.4 and 5.7.1).

Definition 5.7.14 (Strongly convex functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function. Then we say that $\mathscr{L}$ is a strongly convex function (we say that $\mathscr{L}$ is strongly convex) if and only if there exists $c \in (0, \infty)$ such that $\mathscr{L}$ is c-generalized convex (cf.*

Definition 5.7.13).

Lemma 5.7.15 (Equivalence for generalized convex functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function. Then the following two statements are equivalent:*

(i) It holds that $\mathscr{L}$ is c-generalized convex (cf. Definition 5.7.13).

(ii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$\mathscr{L}(tv + (1-t)w) \leq t\mathscr{L}(v) + (1-t)\mathscr{L}(w) - \tfrac{c}{2}\big[t(1-t)\|v-w\|_2^2\big] \tag{5.137}$$

(cf. Definition 3.3.4).

(iii) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$\mathscr{L}(\theta + tv) \leq \mathscr{L}(\theta) + t(\mathscr{L}(\theta+v) - \mathscr{L}(\theta)) - \tfrac{c}{2}\big[t(1-t)\|v\|_2^2\big] \tag{5.138}$$

(cf. Definition 3.3.4).

(iv) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$t\big(\mathscr{L}(\theta+v) - \mathscr{L}(\theta+tv)\big) - (1-t)\big(\mathscr{L}(\theta+tv) - \mathscr{L}(\theta)\big) \geq \tfrac{c}{2}\big[t(1-t)\|v-w\|_2^2\big] \tag{5.139}$$

(cf. Definition 3.3.4).

Proof of Lemma 5.7.15. C8.1Observe that (5.117) and (5.136) imply that $\mathscr{L}$ is c-generalized convex if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$\mathscr{L}(tv+(1-t)w) - \tfrac{c}{2}\|tv+(1-t)w\|_2^2 \leq t\big(\mathscr{L}(v) - \tfrac{c}{2}\|v\|_2^2\big) + (1-t)\big(\mathscr{L}(w) - \tfrac{c}{2}\|w\|_2^2\big) \tag{5.140}$$

(cf. Definitions 3.3.4 and 5.7.13). C5.1Hence, we obtain that $\mathscr{L}$ is c-generalized convex if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$\begin{aligned}\mathscr{L}(tv + (1-t)w) \leq{}& t\mathscr{L}(v) + (1-t)\mathscr{L}(w) \\ &- \tfrac{c}{2}\big(t\|v\|_2^2 + (1-t)\|w\|_2^2 - \|tv + (1-t)w\|_2^2\big).\end{aligned} \tag{5.141}$$

C9.1Moreover, note that the fact that for all $t \in (0,1)$ it holds that

$$(1-t) - (1-t)^2 = 1 - t - t^2 + 2t - 1 = t(1-t) \tag{5.142}$$

shows that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ it holds that

$$\begin{aligned} &t\|v\|_2^2 + (1-t)\|w\|_2^2 - \|tv + (1-t)w\|_2^2 \\ &= t\|v\|_2^2 + (1-t)\|w\|_2^2 - \big(t^2\|v\|_2^2 + (1-t)^2\|w\|_2^2 + 2t(1-t)\langle v, w\rangle\big) \\ &= (t-t^2)\|v\|_2^2 + (1-t-(1-t)^2)\|w\|_2^2 - 2t(1-t)\langle v, w\rangle \\ &= t(1-t)\big(\|v\|_2^2 + \|w\|_2^2 - 2\langle v, w\rangle\big) \\ &= t(1-t)\|v-w\|_2^2. \end{aligned} \tag{5.143}$$

(cf. Definition 1.4.7). C3.2Combining this and (5.141) demonstrates that $\mathscr{L}$ is c-generalized convex if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ that

$$\mathscr{L}(tv + (1-t)w) \leq t\mathscr{L}(v) + (1-t)\mathscr{L}(w) - \tfrac{c}{2}\big[t(1-t)\|v-w\|_2^2\big]. \tag{5.144}$$

C5.2This establishes that ((i) $\leftrightarrow$ (ii)). C9.2Furthermore, observe that (5.137) proves that ((ii) $\leftrightarrow$ (iii)) and that ((iii) $\leftrightarrow$ (iv)). The proof of Lemma 5.7.15 is thus complete. □

Proposition 5.7.16 (Equivalence for differentiable generalized-convex functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable. Then the following three statements are equivalent:*

(i) It holds that $\mathscr{L}$ is c-generalized-convex (cf. Definition 5.7.13).

(ii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \langle (\nabla\mathscr{L})(w), v-w\rangle + \tfrac{c}{2}\|v-w\|_2^2 \tag{5.145}$$

(cf. Definitions 1.4.7 and 3.3.4).

(iii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle (\nabla\mathscr{L})(v) - (\nabla\mathscr{L})(w), v-w\rangle \geq c\|v-w\|_2^2 \tag{5.146}$$

(cf. Definitions 1.4.7 and 3.3.4).

Proof of Proposition 5.7.16. We first prove that ((i) $\to$ (ii)). For this assume that $\mathscr{L}$ is c-generalized convex. C8.1Note that the assumption that $\mathscr{L}$ is c-generalized convex and Lemma 5.7.15 ensure that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ it holds that

$$\mathscr{L}(w + t(v-w)) \leq \mathscr{L}(w) + t(\mathscr{L}(v) - \mathscr{L}(w)) - \tfrac{c}{2}\big[t(1-t)\|w-v\|_2^2\big]. \tag{5.147}$$

(cf. Definitions 3.3.4 and 5.7.13). C5.1Hence, we obtain that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0,1)$ it holds that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \frac{\mathscr{L}(w + t(v-w)) - \mathscr{L}(w)}{t} + \tfrac{c}{2}\big[(1-t)\|v-w\|_2^2\big] \tag{5.148}$$

C3.2Combining this and the assumption that $\mathscr{L}$ is differentiable implies that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \mathscr{L}(v) &\geq \mathscr{L}(w) + \limsup_{t \to 0} \left(\frac{\mathscr{L}(w + t(v-w)) - \mathscr{L}(w)}{t} + \tfrac{c}{2}\left[(1-t)\|v-w\|_2^2\right] \right) \\ &= \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle + \tfrac{c}{2}\|v-w\|_2^2. \end{aligned} \tag{5.149}$$

(cf. Definition 1.4.7). This proves that ((i) → (ii)).

In the next step we prove that ((ii) → (iii)). For this assume that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathscr{L}(v) \geq \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle + \tfrac{c}{2}\|v-w\|_2^2. \tag{5.150}$$

C8.1Observe that (5.150) shows that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \mathscr{L}(v) + \mathscr{L}(w) &\geq \mathscr{L}(w) + \langle (\nabla \mathscr{L})(w), v - w \rangle + \tfrac{c}{2}\|v-w\|_2^2 \\ &\quad + \mathscr{L}(v) + \langle (\nabla \mathscr{L})(v), w - v \rangle + \tfrac{c}{2}\|w-v\|_2^2 \\ &= \mathscr{L}(v) + \mathscr{L}(w) - \langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle + c\|w-v\|_2^2. \end{aligned} \tag{5.151}$$

C5.2This demonstrates that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \geq c\|v-w\|_2^2. \tag{5.152}$$

This proves that ((ii) → (iii)).

In the next step we prove that ((iii) → (i)). For this assume that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \geq c\|v-w\|_2^2. \tag{5.153}$$

C8.1Note that (5.153) establishes that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\alpha, \beta \in \mathbb{R}$ with $\alpha > \beta$ it holds that

$$\begin{aligned} &\langle (\nabla \mathscr{L})(\theta + \alpha v) - (\nabla \mathscr{L})(\theta + \beta v), v \rangle \\ &= (\alpha - \beta)^{-1} \langle (\nabla \mathscr{L})(\theta + \alpha v) - (\nabla \mathscr{L})(\theta + \beta v), (\alpha - \beta) v \rangle \\ &\geq (\alpha - \beta)^{-1} c \|(\alpha - \beta) v\|_2^2 = (\alpha - \beta) c \|v\|_2^2. \end{aligned} \tag{5.154}$$

C3.2Combining this and the fundamental theorem of calculus proves that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$,

$t \in (0,1)$ it holds that

$$\begin{aligned}
&t\big(\mathscr{L}(\theta+v)-\mathscr{L}(\theta+tv)\big)-(1-t)\big(\mathscr{L}(\theta+tv)-\mathscr{L}(\theta)\big)\\
&= t\left(\int_t^1 \langle(\nabla\mathscr{L})(\theta+sv), v\rangle\, ds\right) - (1-t)\left(\int_0^t \langle(\nabla\mathscr{L})(\theta+sv), v\rangle\, ds\right)\\
&= t(1-t)\left(\int_0^1 \langle(\nabla\mathscr{L})(\theta+(t+s(1-t))v), v\rangle\, ds\right)\\
&\quad - (1-t)t\left(\int_0^1 \langle(\nabla\mathscr{L})(\theta+stv), v\rangle\, ds\right) \qquad (5.155)\\
&= t(1-t)\left(\int_0^1 \langle(\nabla\mathscr{L})(\theta+(t+s(1-t))v) - (\nabla\mathscr{L})(\theta+stv), v\rangle\, ds\right)\\
&\geq t(1-t)\left(\int_0^1 (t+s-2st)c\|v\|_2^2\, ds\right)\\
&= t(1-t)(t+\tfrac{1}{2}-t)c\|v\|_2^2 = \tfrac{c}{2}\big[t(1-t)\|v\|_2^2\big]
\end{aligned}$$

C3.1This and Lemma 5.7.15 ensure that $\mathscr{L}$ is c-generalized convex. This proves that ((iii) $\to$ (i)). The proof of Proposition 5.7.16 is thus complete. □

Corollary 5.7.17 (Equivalence for differentiable generalized-convex functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable. Then the following three statements are equivalent:*

(i) It holds that $\mathscr{L}$ is c-generalized-convex (cf. Definition 5.7.13).

(ii) It holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $(\nabla\mathscr{L})(\theta) - c\theta$ is a subgradient of $\mathbb{R}^{\mathfrak{d}} \ni v \mapsto \mathscr{L}(v) - \frac{c}{2}\|v\|_2^2 \in \mathbb{R}$ at θ (cf. Definitions 3.3.4 and 5.7.11).

(iii) It holds that $\nabla\mathscr{L}$ is c-monotonically increasing (cf. Definition 5.7.8).

(iv) It holds that $-\nabla\mathscr{L}$ is $(-c)$-monotonically decreasing (cf. Definition 5.7.9).

Proof of Corollary 5.7.17. C8.1Observe that Lemma 5.7.10, Lemma 5.7.12, Lemma 5.7.15, and (5.133) imply that ((i) $\leftrightarrow$ (ii)), ((ii) $\leftrightarrow$ (iii)), ((iii) $\leftrightarrow$ (iv)), and ((iv) $\leftrightarrow$ (i)). The proof of Corollary 5.7.17 is thus complete. □

5.7.5 Coercivity

Definition 5.7.18 (Coercivity-type conditions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $c \in (0,\infty)$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, and let $\mathscr{L}\colon O \to \mathbb{R}$ be a function. Then we say that $\mathscr{L}$ satisfies a coercivity-type condition with coercivity constant c at ϑ if and only if*

(i) it holds that $\mathscr{L}$ is differentiable and

(ii) it holds for all $\theta \in O$ that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \tag{5.156}$$

(cf. Definitions 1.4.7 and 3.3.4).

Definition 5.7.19 (Coercive-type functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a function. Then we say that $\mathscr{L}$ is a coercive-type function if and only if there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $c \in (0, \infty)$ such that it holds that $\mathscr{L}$ satisfies a coercivity-type condition at ϑ with coercivity constant c (cf. Definition 5.7.18).*

Corollary 5.7.20 (Strongly convex functions are coercive)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable, assume that $\mathscr{L}$ is c-generalized convex, and assume that ϑ is a critical point of $\mathscr{L}$ (cf. Definitions 5.6.5 and 5.7.13). Then it holds that $\mathscr{L}$ satisfies a coercivity-type condition at ϑ with coercivity constant c (cf. Definition 5.7.18).*

Proof of Corollary 5.7.20. C8.1Note that Proposition 5.7.16 shows that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle \geq c \|v - w\|_2^2. \tag{5.157}$$

(cf. Definitions 1.4.7 and 3.3.4). C3.2Combining this and the fact that $(\nabla \mathscr{L})(\vartheta) = 0$ demonstrates that it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle = \langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) - (\nabla \mathscr{L})(\vartheta) \rangle \geq c \|\theta - \vartheta\|_2^2. \tag{5.158}$$

C3.1This and (5.156) establish that $\mathscr{L}$ satisfies a coercivity-type condition at ϑ with coercivity constant c (cf. Definition 5.7.18). The proof of Corollary 5.7.20 is thus complete. □

Corollary 5.7.21. *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be continuously differentiable and strongly convex, and assume that ϑ is a critical point of $\mathscr{L}$ (cf. Definitions 5.6.5 and 5.7.14). Then it holds that $\mathscr{L}$ is a coercive-type function (cf. Definition 5.7.19).*

Proof of Corollary 5.7.21. C8.1Observe that Corollary 5.7.20 proves that $\mathscr{L}$ is a coercive-type function (cf. Definition 5.7.19). The proof of Corollary 5.7.21 is thus complete. □

Lemma 5.7.22 (A sufficient condition for a local minimum point). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \tag{5.159}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) it holds for all $\theta \in \mathbb{B}$ that $\mathscr{L}(\theta) - \mathscr{L}(\vartheta) \geq \frac{c}{2} \|\theta - \vartheta\|_2^2$,

(ii) it holds that $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w \in \mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$, and

(iii) it holds that $(\nabla \mathscr{L})(\vartheta) = 0$.

Proof of Lemma 5.7.22. Throughout this proof, let B be the set given by

$$B = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 < r\}. \tag{5.160}$$

Note that (5.159) implies that for all $v \in \mathbb{R}^{\mathfrak{d}}$ with $\|v\|_2 \leq r$ it holds that

$$\langle (\nabla \mathscr{L})(\vartheta + v), v \rangle \geq c \|v\|_2^2. \tag{5.161}$$

The fundamental theorem of calculus hence demonstrates that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \mathscr{L}(\theta) - \mathscr{L}(\vartheta) &= \big[\mathscr{L}(\vartheta + t(\theta - \vartheta))\big]_{t=0}^{t=1} \\ &= \int_0^1 \mathscr{L}'(\vartheta + t(\theta - \vartheta))(\theta - \vartheta) \, \mathrm{d}t \\ &= \int_0^1 \langle (\nabla \mathscr{L})(\vartheta + t(\theta - \vartheta)), t(\theta - \vartheta) \rangle \frac{1}{t} \, \mathrm{d}t \\ &\geq \int_0^1 c \|t(\theta - \vartheta)\|_2^2 \frac{1}{t} \, \mathrm{d}t = c \|\theta - \vartheta\|_2^2 \left[\int_0^1 t \, \mathrm{d}t \right] = \tfrac{c}{2} \|\theta - \vartheta\|_2^2. \end{aligned} \tag{5.162}$$

This proves item (i). Next observe that (5.162) ensures that for all $\theta \in \mathbb{B} \backslash \{\vartheta\}$ it holds that

$$\mathscr{L}(\theta) \geq \mathscr{L}(\vartheta) + \tfrac{c}{2} \|\theta - \vartheta\|_2^2 > \mathscr{L}(\vartheta). \tag{5.163}$$

Hence, we obtain for all $\theta \in \mathbb{B} \backslash \{\vartheta\}$ that

$$\inf_{w \in \mathbb{B}} \mathscr{L}(w) = \mathscr{L}(\vartheta) < \mathscr{L}(\theta). \tag{5.164}$$

This establishes item (ii). It thus remains thus remains to prove item (iii). For this observe that item (ii) ensures that

$$\{\theta \in B \colon \mathscr{L}(\theta) = \inf_{w \in B} \mathscr{L}(w)\} = \{\vartheta\}. \tag{5.165}$$

Combining this, the fact that B is open, and Lemma 5.6.7 (applied with $\mathfrak{d} \curvearrowleft \mathfrak{d}$, $O \curvearrowleft B$, $\vartheta \curvearrowleft \vartheta$, $\mathscr{L} \curvearrowleft \mathscr{L}|_B$ in the notation of Lemma 5.6.7) assures that $(\nabla \mathscr{L})(\vartheta) = 0$. This establishes item (iii). The proof of Lemma 5.7.22 is thus complete. □

Example 5.7.23. *Let* $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\kappa, \lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}} \in (0, \infty)$ *satisfy* $\kappa = \min\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$, *and let* $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$\mathscr{L}(\theta) = \tfrac{1}{2}\left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2\right]. \tag{5.166}$$

Then

(i) it holds that $\mathscr{L}$ *is* κ*-generalized convex,*

(ii) it holds that $\mathscr{L}$ *is strongly convex,*

(iii) it holds that $\mathscr{L}$ *satisfies a coercivity-type condition at* ϑ *with coercivity constant* κ*, and*

(iv) it holds that $\mathscr{L}$ *is a coercive-type function*

(cf. Definitions 5.7.13, 5.7.14, 5.7.18, and 5.7.19).

Proof for Example 5.7.23. C8.1Note that (6.295) ensures that for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$(\nabla \mathscr{L})(\theta) = (\lambda_1(\theta_1 - \vartheta_1), \ldots, \lambda_{\mathfrak{d}}(\theta_{\mathfrak{d}} - \vartheta_{\mathfrak{d}})). \tag{5.167}$$

C5.1Hence, we obtain that for all $v = (v_1, \ldots, v_{\mathfrak{d}}), w = (w_1, \ldots, w_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \langle (\nabla \mathscr{L})(v) - (\nabla \mathscr{L})(w), v - w \rangle &= \sum_{i=1}^{\mathfrak{d}} \lambda_i (v_i - w_i)(v_i - w_i) \\ &\geq \kappa \sum_{i=1}^{\mathfrak{d}} (v_i - w_i)^2 = \kappa \|v - w\|_2^2 \end{aligned} \tag{5.168}$$

(cf. Definitions 1.4.7 and 3.3.4). C1.1Proposition 5.7.16 hence implies that $\mathscr{L}$ is κ-generalized convex (cf. Definition 5.7.13). This establishes item (i). C8.1Observe that item (i) and the fact that $(\nabla \mathscr{L})(\vartheta) = 0$ establish items (ii), (iii), and (iv). The proof for Example 5.7.23 is thus complete. □

5.8 Lyapunov-type functions for GFs

5.8.1 Gronwall differential inequalities

The following lemma, Lemma 5.8.1 below, is referred to as a Gronwall inequality in the literature (cf., for example, Henry [202, Chapter 7]). Gronwall inequalities are powerful tools to study dynamical systems and, especially, solutions of ODEs.

Lemma 5.8.1 (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*

$$\epsilon'(t) \leq \alpha\epsilon(t) + \beta(t). \tag{5.169}$$

Then it holds for all $t \in [0,T]$ that

$$\epsilon(t) \leq e^{\alpha t}\epsilon(0) + \int_0^t e^{\alpha(t-s)}\beta(s)\,\mathrm{d}s. \tag{5.170}$$

Proof of Lemma 5.8.1. Throughout this proof, let $v\colon [0,T] \to \mathbb{R}$ satisfy for all $t \in [0,T]$ that

$$v(t) = e^{\alpha t}\left[\int_0^t e^{-\alpha s}\beta(s)\,\mathrm{d}s\right] \tag{5.171}$$

and let $u\colon [0,T] \to \mathbb{R}$ satisfy for all $t \in [0,T]$ that

$$u(t) = [\epsilon(t) - v(t)]e^{-\alpha t}. \tag{5.172}$$

Note 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

$$v'(t) = \alpha e^{\alpha t}\left[\int_0^t e^{-\alpha s}\beta(s)\,\mathrm{d}s\right] + e^{\alpha t}\left[e^{-\alpha t}\beta(t)\right] = \alpha v(t) + \beta(t). \tag{5.173}$$

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

$$\begin{aligned} u'(t) &= [\epsilon'(t) - v'(t)]e^{-\alpha t} - [\epsilon(t) - v(t)]\alpha e^{-\alpha t} \\ &= [\epsilon'(t) - v'(t) - \alpha\epsilon(t) + \alpha v(t)]e^{-\alpha t} \\ &= [\epsilon'(t) - \alpha v(t) - \beta(t) - \alpha\epsilon(t) + \alpha v(t)]e^{-\alpha t} \\ &= [\epsilon'(t) - \beta(t) - \alpha\epsilon(t)]e^{-\alpha t}. \end{aligned} \tag{5.174}$$

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

$$u'(t) \leq [\alpha\epsilon(t) + \beta(t) - \beta(t) - \alpha\epsilon(t)]e^{-\alpha t} = 0. \tag{5.175}$$

This and the fundamental theorem of calculus imply that for all $t \in [0,T]$ it holds that

$$u(t) = u(0) + \int_0^t u'(s)\,\mathrm{d}s \leq u(0) + \int_0^t 0\,\mathrm{d}s = u(0) = \epsilon(0). \tag{5.176}$$

Combining this, (5.171), and (5.172) shows that for all $t \in [0,T]$ it holds that

$$\epsilon(t) = e^{\alpha t}u(t) + v(t) \leq e^{\alpha t}\epsilon(0) + v(t) = e^{\alpha t}\epsilon(0) + \int_0^t e^{\alpha(t-s)}\beta(s)\,\mathrm{d}s. \tag{5.177}$$

The proof of Lemma 5.8.1 is thus complete. □

5.8.2 Lyapunov-type functions for ODEs

Proposition 5.8.2 (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 $\beta \in C(O,\mathbb{R})$, $\mathscr{G} \in C(O,\mathbb{R}^{\mathfrak{d}})$, $V \in C^1(O,\mathbb{R})$ satisfy for all $\theta \in O$ that*

$$V'(\theta)\mathscr{G}(\theta) = \langle (\nabla V)(\theta), \mathscr{G}(\theta)\rangle \leq \alpha V(\theta) + \beta(\theta), \tag{5.178}$$

and let $\Theta \in C([0,T],O)$ satisfy for all $t \in [0,T]$ that $\Theta_t = \Theta_0 + \int_0^t \mathscr{G}(\Theta_s)\,\mathrm{d}s$ (cf. Definition 1.4.7). Then it holds for all $t \in [0,T]$ that

$$V(\Theta_t) \leq e^{\alpha t} V(\Theta_0) + \int_0^t e^{\alpha(t-s)}\beta(\Theta_s)\,\mathrm{d}s. \tag{5.179}$$

Proof of Proposition 5.8.2. Throughout this proof, let $\epsilon, b \in C([0,T],\mathbb{R})$ satisfy for all $t \in [0,T]$ that

$$\epsilon(t) = V(\Theta_t) \qquad \text{and} \qquad b(t) = \beta(\Theta_t). \tag{5.180}$$

Observe that (5.178), (5.180), the fundamental theorem of calculus, and the chain rule ensure that for all $t \in [0,T]$ it holds that

$$\epsilon'(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}(V(\Theta_t)) = V'(\Theta_t)\dot{\Theta}_t = V'(\Theta_t)\mathscr{G}(\Theta_t) \leq \alpha V(\Theta_t) + \beta(\Theta_t) = \alpha\epsilon(t) + b(t). \tag{5.181}$$

Lemma 5.8.1 and (5.180) hence demonstrate that for all $t \in [0,T]$ it holds that

$$V(\Theta_t) = \epsilon(t) \leq e^{\alpha t}\epsilon(0) + \int_0^t e^{\alpha(t-s)} b(s)\,\mathrm{d}s = e^{\alpha t} V(\Theta_0) + \int_0^t e^{\alpha(t-s)}\beta(\Theta_s)\,\mathrm{d}s. \tag{5.182}$$

The proof of Proposition 5.8.2 is thus complete. □

Corollary 5.8.3. *Let $\mathfrak{d} \in \mathbb{N}$, $T \in (0,\infty)$, $\alpha \in \mathbb{R}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\mathscr{G} \in C(O,\mathbb{R}^{\mathfrak{d}})$, $V \in C^1(O,\mathbb{R})$ satisfy for all $\theta \in O$ that*

$$V'(\theta)\mathscr{G}(\theta) = \langle (\nabla V)(\theta), \mathscr{G}(\theta)\rangle \leq \alpha V(\theta), \tag{5.183}$$

and let $\Theta \in C([0,T],O)$ satisfy for all $t \in [0,T]$ that $\Theta_t = \Theta_0 + \int_0^t \mathscr{G}(\Theta_s)\,\mathrm{d}s$ (cf. Definition 1.4.7). Then it holds for all $t \in [0,T]$ that

$$V(\Theta_t) \leq e^{\alpha t} V(\Theta_0). \tag{5.184}$$

Proof of Corollary 5.8.3. Note that Proposition 5.8.2 and (5.183) show (5.184). The proof of Corollary 5.8.3 is thus complete. □

5.8.3 On Lyapunov-type functions and coercivity-type conditions

Lemma 5.8.4 (Derivative of the standard norm). *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and let $V \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$V(\theta) = \|\theta - \vartheta\|_2^2 \tag{5.185}$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $V \in C^\infty(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and

$$(\nabla V)(\theta) = 2(\theta - \vartheta). \tag{5.186}$$

Proof of Lemma 5.8.4. Throughout this proof, let $\vartheta_1, \ldots, \vartheta_{\mathfrak{d}} \in \mathbb{R}$ satisfy $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}})$. Note that the fact that for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$V(\theta) = \sum_{i=1}^{\mathfrak{d}} (\theta_i - \vartheta_i)^2 \tag{5.187}$$

implies that for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that $V \in C^\infty(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and

$$(\nabla V)(\theta) = \begin{pmatrix} (\frac{\partial V}{\partial \theta_1})(\theta) \\ \vdots \\ (\frac{\partial V}{\partial \theta_{\mathfrak{d}}})(\theta) \end{pmatrix} = \begin{pmatrix} 2(\theta_1 - \vartheta_1) \\ \vdots \\ 2(\theta_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}) \end{pmatrix} = 2(\theta - \vartheta). \tag{5.188}$$

The proof of Lemma 5.8.4 is thus complete. □

In the next result, Corollary 5.8.5 below, we establish an error analysis for GFs in which the objective function satisfies a coercivity-type condition in the sense of Definition 5.7.18.

Corollary 5.8.5 (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 $\mathscr{L} \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2, \tag{5.189}$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 - \int_0^t (\nabla \mathscr{L})(\Theta_s)\, \mathrm{d}s$ (cf. Definitions 1.4.7 and 3.3.4). Then it holds for all $t \in [0, T]$ that

$$\|\Theta_t - \vartheta\|_2 \leq e^{-ct}\|\Theta_0 - \vartheta\|_2. \tag{5.190}$$

Proof of Corollary 5.8.5. Throughout this proof, let $\mathscr{G} \colon O \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in O$ that

$$\mathscr{G}(\theta) = -(\nabla \mathscr{L})(\theta) \tag{5.191}$$

and let $V \colon O \to \mathbb{R}$ satisfy for all $\theta \in O$ that

$$V(\theta) = \|\theta - \vartheta\|_2^2. \tag{5.192}$$

Observe that Lemma 5.8.4 and (5.189) ensure that for all $\theta \in O$ it holds that $V \in C^1(O, \mathbb{R})$ and

$$\begin{aligned} V'(\theta)\mathscr{G}(\theta) &= \langle (\nabla V)(\theta), \mathscr{G}(\theta) \rangle = \langle 2(\theta - \vartheta), \mathscr{G}(\theta) \rangle \\ &= -2\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \leq -2c\|\theta - \vartheta\|_2^2 = -2cV(\theta). \end{aligned} \tag{5.193}$$

Corollary 5.8.3 hence proves that for all $t \in [0, T]$ it holds that

$$\|\Theta_t - \vartheta\|_2^2 = V(\Theta_t) \leq e^{-2ct} V(\Theta_0) = e^{-2ct} \|\Theta_0 - \vartheta\|_2^2. \tag{5.194}$$

The proof of Corollary 5.8.5 is thus complete. □

5.8.4 On a linear growth condition

Lemma 5.8.6 (On a linear growth condition)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $L \in \mathbb{R}$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2 \tag{5.195}$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$\mathscr{L}(\theta) - \mathscr{L}(\vartheta) \leq \tfrac{L}{2}\|\theta - \vartheta\|_2^2. \tag{5.196}$$

Proof of Lemma 5.8.6. Observe that (5.195), the Cauchy-Schwarz inequality, and the fundamental theorem of calculus ensure that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \mathscr{L}(\theta) - \mathscr{L}(\vartheta) &= \big[\mathscr{L}(\vartheta + t(\theta - \vartheta))\big]_{t=0}^{t=1} \\ &= \int_0^1 \mathscr{L}'(\vartheta + t(\theta - \vartheta))(\theta - \vartheta)\, \mathrm{d}t \\ &= \int_0^1 \langle (\nabla \mathscr{L})(\vartheta + t(\theta - \vartheta)), \theta - \vartheta \rangle\, \mathrm{d}t \\ &\leq \int_0^1 \|(\nabla \mathscr{L})(\vartheta + t(\theta - \vartheta))\|_2 \|\theta - \vartheta\|_2\, \mathrm{d}t \\ &\leq \int_0^1 L\|\vartheta + t(\theta - \vartheta) - \vartheta\|_2 \|\theta - \vartheta\|_2\, \mathrm{d}t \\ &= L\|\theta - \vartheta\|_2^2 \left[\int_0^1 t\, \mathrm{d}t\right] = \tfrac{L}{2}\|\theta - \vartheta\|_2^2 \end{aligned} \tag{5.197}$$

(cf. Definition 1.4.7). The proof of Lemma 5.8.6 is thus complete. □

5.9 Optimization through flows of ODEs

5.9.1 Approximation of local minimum points through GFs

Proposition 5.9.1 (Approximation of local minimum points through GFs)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $c, T \in (0,\infty)$, $r \in (0,\infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *satisfy for all* $\theta \in \mathbb{B}$ *that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2, \tag{5.198}$$

and let $\Theta \in C([0,T], \mathbb{R}^{\mathfrak{d}})$ *satisfy for all* $t \in [0,T]$ *that* $\Theta_t = \xi - \int_0^t (\nabla \mathscr{L})(\Theta_s)\,\mathrm{d}s$ *(cf. Definitions 1.4.7 and 3.3.4). Then*

(i) it holds that $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w \in \mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(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*

$$0 \leq \tfrac{c}{2}\|\Theta_t - \vartheta\|_2^2 \leq \mathscr{L}(\Theta_t) - \mathscr{L}(\vartheta). \tag{5.199}$$

Proof of Proposition 5.9.1. Throughout this proof, let $V \colon \mathbb{R}^{\mathfrak{d}} \to [0,\infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $V(\theta) = \|\theta - \vartheta\|_2^2$, let $\epsilon \colon [0,T] \to [0,\infty)$ 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

$$\tau = \inf(\{t \in [0,T] \colon \Theta_t \notin \mathbb{B}\} \cup \{T\}) = \inf\big(\{t \in [0,T] \colon \epsilon(t) > r^2\} \cup \{T\}\big). \tag{5.200}$$

Note that (5.198) and item (ii) in Lemma 5.7.22 establish item (i). Next observe that Lemma 5.8.4 implies that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $V \in C^1(\mathbb{R}^{\mathfrak{d}}, [0,\infty))$ and

$$(\nabla V)(\theta) = 2(\theta - \vartheta). \tag{5.201}$$

Moreover, observe that the fundamental theorem of calculus (see, for instance, Coleman [87, Theorem 3.9]) and the fact that $\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathscr{L})(v) \in \mathbb{R}^{\mathfrak{d}}$ and $\Theta \colon [0,T] \to \mathbb{R}^{\mathfrak{d}}$ 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

$$\tfrac{\mathrm{d}}{\mathrm{d}t}(\Theta_t) = -(\nabla \mathscr{L})(\Theta_t). \tag{5.202}$$

Combining (5.198) and (5.201) hence demonstrates that for all $t \in [0,\tau]$ it holds that $\epsilon \in C^1([0,T],[0,\infty))$ and

$$\begin{aligned} \epsilon'(t) &= \tfrac{\mathrm{d}}{\mathrm{d}t}\big(V(\Theta_t)\big) = V'(\Theta_t)\big(\tfrac{\mathrm{d}}{\mathrm{d}t}(\Theta_t)\big) \\ &= \langle (\nabla V)(\Theta_t), \tfrac{\mathrm{d}}{\mathrm{d}t}(\Theta_t) \rangle \\ &= \langle 2(\Theta_t - \vartheta), -(\nabla \mathscr{L})(\Theta_t) \rangle \\ &= -2\langle (\Theta_t - \vartheta), (\nabla \mathscr{L})(\Theta_t) \rangle \\ &\leq -2c\|\Theta_t - \vartheta\|_2^2 = -2c\epsilon(t). \end{aligned} \tag{5.203}$$

The Gronwall inequality, for example, in Lemma 5.8.1 therefore implies that for all $t \in [0,\tau]$ it holds that

$$\epsilon(t) \leq \epsilon(0) e^{-2ct}. \tag{5.204}$$

Hence, we obtain for all $t \in [0,\tau]$ that

$$\|\Theta_t - \vartheta\|_2 = \sqrt{\epsilon(t)} \leq \sqrt{\epsilon(0)} e^{-ct} = \|\Theta_0 - \vartheta\|_2 e^{-ct} = \|\xi - \vartheta\|_2 e^{-ct}. \tag{5.205}$$

In the next step we prove that

$$\tau > 0. \tag{5.206}$$

In our proof of (5.206) we distinguish between the case $\varepsilon(0) = 0$ and the case $\varepsilon(0) > 0$. We first prove (5.206) in the case

$$\varepsilon(0) = 0. \tag{5.207}$$

Observe that (5.207), the assumption that $r \in (0,\infty]$, and the fact that $\epsilon\colon [0,T] \to [0,\infty)$ is a continuous function show that

$$\tau = \inf\big(\{t \in [0,T]\colon \epsilon(t) > r^2\} \cup \{T\}\big) > 0. \tag{5.208}$$

This establishes (5.206) in the case $\varepsilon(0) = 0$. In the next step we prove (5.206) in the case

$$\varepsilon(0) > 0. \tag{5.209}$$

Note that (5.203) and the assumption that $c \in (0,\infty)$ assure that for all $t \in [0,\tau]$ with $\epsilon(t) > 0$ it holds that

$$\epsilon'(t) \leq -2c\epsilon(t) < 0. \tag{5.210}$$

Combining this with (5.209) shows that

$$\epsilon'(0) < 0. \tag{5.211}$$

The fact that $\epsilon'\colon [0,T] \to [0,\infty)$ is a continuous function and the assumption that $T \in (0,\infty)$ therefore demonstrate that

$$\inf(\{t \in [0,T]\colon \epsilon'(t) > 0\} \cup \{T\}) > 0. \tag{5.212}$$

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 $s < \inf(\{t \in [0,T]\colon \epsilon'(t) > 0\} \cup \{T\})$ it holds that

$$\epsilon(s) = \epsilon(0) + \int_0^s \epsilon'(u)\,\mathrm{d}u \leq \epsilon(0) = \|\xi - \vartheta\|_2^2 \leq r^2. \tag{5.213}$$

Combining this with (5.212) proves that

$$\tau = \inf\big(\{s \in [0,T]\colon \epsilon(s) > r^2\} \cup \{T\}\big) > 0. \tag{5.214}$$

This establishes (5.206) in the case $\varepsilon(0) > 0$. Observe that (5.205), (5.206), and the assumption that $c \in (0, \infty)$ demonstrate that

$$\|\Theta_\tau - \vartheta\|_2 \leq \|\xi - \vartheta\|_2 e^{-c\tau} < r. \tag{5.215}$$

The fact that $\epsilon \colon [0, T] \to [0, \infty)$ is a continuous function, (5.200), and (5.206) hence assure that $\tau = T$. Combining this with (5.205) proves that for all $t \in [0, T]$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq \|\xi - \vartheta\|_2 e^{-ct}. \tag{5.216}$$

This establishes item (ii). It thus remains to prove item (iii). For this observe that (5.198) and item (i) in Lemma 5.7.22 demonstrate that for all $\theta \in \mathbb{B}$ it holds that

$$0 \leq \tfrac{c}{2}\|\theta - \vartheta\|_2^2 \leq \mathscr{L}(\theta) - \mathscr{L}(\vartheta). \tag{5.217}$$

Combining this and item (ii) implies that for all $t \in [0, T]$ it holds that

$$0 \leq \tfrac{c}{2}\|\Theta_t - \vartheta\|_2^2 \leq \mathscr{L}(\Theta_t) - \mathscr{L}(\vartheta) \tag{5.218}$$

This establishes item (iii). The proof of Proposition 5.9.1 is thus complete. □

5.9.2 Existence and uniqueness of solutions of ODEs

Lemma 5.9.2 (Local existence of maximal solution of ODEs)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $T \in (0, \infty)$, let $\|\!|\cdot|\!\| \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty)$ be a norm, and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be locally Lipschitz continuous. Then there exist a unique real number $\tau \in (0, T]$ and a unique continuous function $\Theta \colon [0, \tau) \to \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0, \tau)$ it holds that*

$$\liminf_{s \nearrow \tau} \left[\|\!|\Theta_s|\!\| + \tfrac{1}{(T-s)} \right] = \infty \qquad \text{and} \qquad \Theta_t = \xi + \int_0^t \mathscr{G}(\Theta_s)\, \mathrm{d}s. \tag{5.219}$$

Proof of Lemma 5.9.2. Note that, for instance, Teschl [408, Theorem 2.2 and Corollary 2.16] implies (5.219) (cf., for example, [5, Theorem 7.6] and [230, Theorem 1.1]). The proof of Lemma 5.9.2 is thus complete. □

Lemma 5.9.3 (Local existence of maximal solution of ODEs on an infinite time interval)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $\|\!|\cdot|\!\| \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty)$ be a norm, and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be locally Lipschitz continuous. Then there exist a unique extended real number $\tau \in (0, \infty]$ and a unique continuous function $\Theta \colon [0, \tau) \to \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0, \tau)$ it holds that*

$$\liminf_{s \nearrow \tau} \left[\|\!|\Theta_s|\!\| + s \right] = \infty \qquad \text{and} \qquad \Theta_t = \xi + \int_0^t \mathscr{G}(\Theta_s)\, \mathrm{d}s. \tag{5.220}$$

Proof of Lemma 5.9.3. First, observe that Lemma 5.9.2 implies that there exist unique real numbers $\tau_n \in (0, n]$, $n \in \mathbb{N}$, and unique continuous functions $\Theta^{(n)} \colon [0, \tau_n) \to \mathbb{R}^{\mathfrak{d}}$, $n \in \mathbb{N}$, such that for all $n \in \mathbb{N}$, $t \in [0, \tau_n)$ it holds that

$$\liminf_{s \nearrow \tau_n} \left[\left\| \Theta_s^{(n)} \right\| + \tfrac{1}{(n-s)} \right] = \infty \qquad \text{and} \qquad \Theta_t^{(n)} = \xi + \int_0^t \mathscr{G}(\Theta_s^{(n)}) \, \mathrm{d}s. \tag{5.221}$$

This shows that for all $n \in \mathbb{N}$, $t \in [0, \min\{\tau_{n+1}, n\})$ it holds that

$$\liminf_{s \nearrow \tau_{n+1}} \left[\left\| \Theta_s^{(n+1)} \right\| + \tfrac{1}{(n+1-s)} \right] = \infty \qquad \text{and} \qquad \Theta_t^{(n+1)} = \xi + \int_0^t \mathscr{G}(\Theta_s^{(n+1)}) \, \mathrm{d}s. \tag{5.222}$$

Hence, we obtain that for all $n \in \mathbb{N}$, $t \in [0, \min\{\tau_{n+1}, n\})$ it holds that

$$\liminf_{s \nearrow \min\{\tau_{n+1}, n\}} \left[\left\| \Theta_s^{(n+1)} \right\| + \tfrac{1}{(n-s)} \right] = \infty \tag{5.223}$$

$$\text{and} \qquad \Theta_t^{(n+1)} = \xi + \int_0^t \mathscr{G}(\Theta_s^{(n+1)}) \, \mathrm{d}s. \tag{5.224}$$

Combining this with (5.221) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\tau_n = \min\{\tau_{n+1}, n\} \qquad \text{and} \qquad \Theta^{(n)} = \Theta^{(n+1)}|_{[0, \min\{\tau_{n+1}, n\})}. \tag{5.225}$$

Therefore, we obtain that for all $n \in \mathbb{N}$ it holds that

$$\tau_n \leq \tau_{n+1} \qquad \text{and} \qquad \Theta^{(n)} = \Theta^{(n+1)}|_{[0, \tau_n)}. \tag{5.226}$$

Next let $\mathfrak{t} \in (0, \infty]$ be the extended real number given by

$$\mathfrak{t} = \lim_{n \to \infty} \tau_n \tag{5.227}$$

and let $\boldsymbol{\Theta} \colon [0, \mathfrak{t}) \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $t \in [0, \tau_n)$ that

$$\boldsymbol{\Theta}_t = \Theta_t^{(n)}. \tag{5.228}$$

Observe that for all $t \in [0, \mathfrak{t})$ there exists $n \in \mathbb{N}$ such that $t \in [0, \tau_n)$. This, (5.221), and (5.226) assure that for all $t \in [0, \mathfrak{t})$ it holds that $\boldsymbol{\Theta} \in C([0, \mathfrak{t}), \mathbb{R}^{\mathfrak{d}})$ and

$$\boldsymbol{\Theta}_t = \xi + \int_0^t \mathscr{G}(\boldsymbol{\Theta}_s) \, \mathrm{d}s. \tag{5.229}$$

In addition, note that (5.225) ensures that for all $n \in \mathbb{N}$, $k \in \mathbb{N} \cap [n, \infty)$ it holds that

$$\min\{\tau_{k+1}, n\} = \min\{\tau_{k+1}, k, n\} = \min\{\min\{\tau_{k+1}, k\}, n\} = \min\{\tau_k, n\}. \tag{5.230}$$

This shows that for all $n \in \mathbb{N}$, $k \in \mathbb{N} \cap (n, \infty)$ it holds that $\min\{\tau_k, n\} = \min\{\tau_{k-1}, n\}$. Hence, we obtain that for all $n \in \mathbb{N}$, $k \in \mathbb{N} \cap (n, \infty)$ it holds that

$$\min\{\tau_k, n\} = \min\{\tau_{k-1}, n\} = \ldots = \min\{\tau_{n+1}, n\} = \min\{\tau_n, n\} = \tau_n. \tag{5.231}$$

Combining this with the fact that $(\tau_n)_{n\in\mathbb{N}} \subseteq [0, \infty)$ is a non-decreasing sequence implies that for all $n \in \mathbb{N}$ it holds that

$$\min\{\mathfrak{t}, n\} = \min\Big\{\lim_{k\to\infty} \tau_k, n\Big\} = \lim_{k\to\infty} \big(\min\{\tau_k, n\}\big) = \lim_{k\to\infty} \tau_n = \tau_n. \tag{5.232}$$

Therefore, we obtain that for all $n \in \mathbb{N}$ with $\mathfrak{t} < n$ it holds that

$$\tau_n = \min\{\mathfrak{t}, n\} = \mathfrak{t}. \tag{5.233}$$

This, (5.221), and (5.228) demonstrate that for all $n \in \mathbb{N}$ with $\mathfrak{t} < n$ it holds that

$$\begin{aligned}
\liminf_{s\nearrow\mathfrak{t}} \|\!|\boldsymbol{\Theta}_s|\!\| &= \liminf_{s\nearrow\tau_n} \|\!|\boldsymbol{\Theta}_s|\!\| = \liminf_{s\nearrow\tau_n} \|\!|\Theta^{(n)}_s|\!\| \\
&= -\tfrac{1}{(n-\mathfrak{t})} + \liminf_{s\nearrow\tau_n} \Big[\|\!|\Theta^{(n)}_s|\!\| + \tfrac{1}{(n-\mathfrak{t})} \Big] \\
&= -\tfrac{1}{(n-\mathfrak{t})} + \liminf_{s\nearrow\tau_n} \Big[\|\!|\Theta^{(n)}_s|\!\| + \tfrac{1}{(n-s)} \Big] = \infty.
\end{aligned} \tag{5.234}$$

Therefore, we obtain that

$$\liminf_{s\nearrow\mathfrak{t}} \big[\|\!|\boldsymbol{\Theta}_s|\!\| + s \big] = \infty. \tag{5.235}$$

Next note that for all $\hat{\mathfrak{t}} \in (0, \infty]$, $\hat{\boldsymbol{\Theta}} \in C([0, \hat{\mathfrak{t}}), \mathbb{R}^{\mathfrak{d}})$, $n \in \mathbb{N}$, $t \in [0, \min\{\hat{\mathfrak{t}}, n\})$ with $\liminf_{s\nearrow\hat{\mathfrak{t}}} [\|\!|\hat{\boldsymbol{\Theta}}_s|\!\| + s] = \infty$ and $\forall\, s \in [0, \hat{\mathfrak{t}})\colon \hat{\boldsymbol{\Theta}}_s = \xi + \int_0^s \mathscr{G}(\hat{\boldsymbol{\Theta}}_u)\, \mathrm{d}u$ it holds that

$$\liminf_{s\nearrow\min\{\hat{\mathfrak{t}},n\}} \Big[\|\!|\hat{\boldsymbol{\Theta}}_s|\!\| + \tfrac{1}{(n-s)} \Big] = \infty \qquad \text{and} \qquad \hat{\boldsymbol{\Theta}}_t = \xi + \int_0^t \mathscr{G}(\hat{\boldsymbol{\Theta}}_s)\, \mathrm{d}s. \tag{5.236}$$

This and (5.221) prove that for all $\hat{\mathfrak{t}} \in (0, \infty]$, $\hat{\boldsymbol{\Theta}} \in C([0, \hat{\mathfrak{t}}), \mathbb{R}^{\mathfrak{d}})$, $n \in \mathbb{N}$ with $\liminf_{t\nearrow\hat{\mathfrak{t}}} [\|\!|\hat{\boldsymbol{\Theta}}_t|\!\| + t] = \infty$ and $\forall\, t \in [0, \hat{\mathfrak{t}})\colon \hat{\boldsymbol{\Theta}}_t = \xi + \int_0^t \mathscr{G}(\hat{\boldsymbol{\Theta}}_s)\, \mathrm{d}s$ it holds that

$$\min\{\hat{\mathfrak{t}}, n\} = \tau_n \qquad \text{and} \qquad \hat{\boldsymbol{\Theta}}|_{[0,\tau_n)} = \Theta^{(n)}. \tag{5.237}$$

Combining (5.229) and (5.235) hence assures that for all $\hat{\mathfrak{t}} \in (0, \infty]$, $\hat{\boldsymbol{\Theta}} \in C([0, \hat{\mathfrak{t}}), \mathbb{R}^{\mathfrak{d}})$, $n \in \mathbb{N}$ with $\liminf_{t\nearrow\hat{\mathfrak{t}}} [\|\!|\hat{\boldsymbol{\Theta}}_t|\!\| + t] = \infty$ and $\forall\, t \in [0, \hat{\mathfrak{t}})\colon \hat{\boldsymbol{\Theta}}_t = \xi + \int_0^t \mathscr{G}(\hat{\boldsymbol{\Theta}}_s)\, \mathrm{d}s$ it holds that

$$\min\{\hat{\mathfrak{t}}, n\} = \tau_n = \min\{\mathfrak{t}, n\} \qquad \text{and} \qquad \hat{\boldsymbol{\Theta}}|_{[0,\tau_n)} = \Theta^{(n)} = \boldsymbol{\Theta}|_{[0,\tau_n)}. \tag{5.238}$$

This and (5.227) show that for all $\hat{\mathfrak{t}} \in (0, \infty]$, $\hat{\boldsymbol{\Theta}} \in C([0, \hat{\mathfrak{t}}), \mathbb{R}^{\mathfrak{d}})$ with $\liminf_{t\nearrow\hat{\mathfrak{t}}} [\|\!|\hat{\boldsymbol{\Theta}}_t|\!\| + t] = \infty$ and $\forall\, t \in [0, \hat{\mathfrak{t}})\colon \hat{\boldsymbol{\Theta}}_t = \xi + \int_0^t \mathscr{G}(\hat{\boldsymbol{\Theta}}_s)\, \mathrm{d}s$ it holds that

$$\hat{\mathfrak{t}} = \mathfrak{t} \qquad \text{and} \qquad \hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}. \tag{5.239}$$

Combining this, (5.229), and (5.235) completes the proof of Lemma 5.9.3. □

5.9.3 Approximation of local minimum points through GFs revisited

Theorem 5.9.4 (Approximation of local minimum points through GFs revisited)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0,\infty)$, $r \in (0,\infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathscr{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2 \tag{5.240}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) there exists a unique continuous function $\Theta \colon [0,\infty) \to \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0,\infty)$ it holds that

$$\Theta_t = \xi - \int_0^t (\nabla \mathscr{L})(\Theta_s)\,\mathrm{d}s, \tag{5.241}$$

(ii) it holds that $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w \in \mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(iii) it holds for all $t \in [0,\infty)$ that $\|\Theta_t - \vartheta\|_2 \leq e^{-ct}\|\xi - \vartheta\|_2$, and

(iv) it holds for all $t \in [0,\infty)$ that

$$0 \leq \tfrac{c}{2}\|\Theta_t - \vartheta\|_2^2 \leq \mathscr{L}(\Theta_t) - \mathscr{L}(\vartheta). \tag{5.242}$$

Proof of Theorem 5.9.4. First, observe that the assumption that $\mathscr{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ ensures that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto -(\nabla \mathscr{L})(\theta) \in \mathbb{R}^{\mathfrak{d}} \tag{5.243}$$

is continuously differentiable. The fundamental theorem of calculus hence implies that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto -(\nabla \mathscr{L})(\theta) \in \mathbb{R}^{\mathfrak{d}} \tag{5.244}$$

is locally Lipschitz continuous. Combining this with Lemma 5.9.3 (applied with $\mathscr{G} \curvearrowright (\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto -(\nabla \mathscr{L})(\theta) \in \mathbb{R}^{\mathfrak{d}})$ in the notation of Lemma 5.9.3) proves that there exists a unique extended real number $\tau \in (0,\infty]$ and a unique continuous function $\Theta \colon [0,\tau) \to \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0,\tau)$ it holds that

$$\liminf_{s \nearrow \tau} \big[\|\Theta_s\|_2 + s\big] = \infty \qquad \text{and} \qquad \Theta_t = \xi - \int_0^t (\nabla \mathscr{L})(\Theta_s)\,\mathrm{d}s. \tag{5.245}$$

Next observe that Proposition 5.9.1 proves that for all $t \in [0,\tau)$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq e^{-ct}\|\xi - \vartheta\|_2. \tag{5.246}$$

This implies that

$$\begin{aligned}\liminf_{s\nearrow\tau}\|\Theta_s\|_2 &\leq \left[\liminf_{s\nearrow\tau}\|\Theta_s-\vartheta\|_2\right]+\|\vartheta\|_2\\ &\leq \left[\liminf_{s\nearrow\tau} e^{-cs}\|\xi-\vartheta\|_2\right]+\|\vartheta\|_2 \leq \|\xi-\vartheta\|_2+\|\vartheta\|_2<\infty.\end{aligned} \tag{5.247}$$

This and (5.245) demonstrate that

$$\tau=\infty. \tag{5.248}$$

This and (5.245) prove item (i). Moreover, note that Proposition 5.9.1 and item (i) establish items (ii), (iii), and (iv). The proof of Theorem 5.9.4 is thus complete. □

5.9.4 Approximation error with respect to the objective function

Corollary 5.9.5 (Approximation error with respect to the objective function)**.** *Let* $\mathfrak{d}\in\mathbb{N}$, $c,L\in(0,\infty)$, $r\in(0,\infty]$, $\vartheta\in\mathbb{R}^{\mathfrak{d}}$, $\mathbb{B}=\{w\in\mathbb{R}^{\mathfrak{d}}\colon\|w-\vartheta\|_2\leq r\}$, $\xi\in\mathbb{B}$, $\mathscr{L}\in C^2(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$ *satisfy for all* $\theta\in\mathbb{B}$ *that*

$$\langle\theta-\vartheta,(\nabla\mathscr{L})(\theta)\rangle\geq c\|\theta-\vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla\mathscr{L})(\theta)\|_2\leq L\|\theta-\vartheta\|_2 \tag{5.249}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) there exists a unique continuous function $\Theta\colon[0,\infty)\to\mathbb{R}^{\mathfrak{d}}$ *such that for all* $t\in[0,\infty)$ *it holds that*

$$\Theta_t=\xi-\int_0^t(\nabla\mathscr{L})(\Theta_s)\,\mathrm{d}s, \tag{5.250}$$

(ii) it holds that $\{\theta\in\mathbb{B}\colon\mathscr{L}(\theta)=\inf_{w\in\mathbb{B}}\mathscr{L}(w)\}=\{\vartheta\}$,

(iii) it holds for all $t\in[0,\infty)$ *that* $\|\Theta_t-\vartheta\|_2\leq e^{-ct}\|\xi-\vartheta\|_2$, *and*

(iv) it holds for all $t\in[0,\infty)$ *that*

$$0\leq\tfrac{c}{2}\|\Theta_t-\vartheta\|_2^2\leq\mathscr{L}(\Theta_t)-\mathscr{L}(\vartheta)\leq\tfrac{L}{2}\|\Theta_t-\vartheta\|_2^2\leq\tfrac{L}{2}e^{-2ct}\|\xi-\vartheta\|_2^2. \tag{5.251}$$

Proof of Corollary 5.9.5. Theorem 5.9.4 and Lemma 5.8.6 establish items (i), (ii), (iii), and (iv). The proof of Corollary 5.9.5 is thus complete. □

Chapter 6

Deterministic gradient descent (GD) optimization methods

This chapter reviews and studies deterministic GD-type optimization methods such as the classical plain-vanilla GD optimization method (see Section 6.1 below) as well as more sophisticated GD-type optimization methods including GD optimization methods with momenta (cf. Sections 6.3, 6.4, and 6.8 below) and GD optimization methods with adaptive modifications of the learning rates (cf. Sections 6.5, 6.6, 6.7, and 6.8 below).

There are several other outstanding reviews on gradient based optimization methods in the literature; cf., for instance, the books [9, Chapter 5], [53, Chapter 9], [58, Chapter 3], [170, Sections 4.3 and 5.9 and Chapter 8], [316], and [387, Chapter 14] and the references therein and, for example, the survey articles [33, 49, 127, 368, 400] and the references therein.

6.1 GD optimization

In this section we review and study the classical plain-vanilla GD optimization method (cf., for example, [316, Section 1.2.3], [53, Section 9.3], and [58, Chapter 3]). 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 (see, for example, Theorem 5.9.4 in Chapter 5 above)

Definition 6.1.1 (GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say*

that Θ is the GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ if and only if it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla\mathscr{L})(\Theta_{n-1}). \tag{6.1}$$

Algorithm 6.1.2: GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ (cf. Definition 6.1.1)

1: **Initialization:** $\Theta \leftarrow \xi$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \Theta \leftarrow \Theta - \gamma_n(\nabla\mathscr{L})(\Theta)$
4: **return** Θ

Exercise 6.1.1. Let $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$ satisfy $\xi = (1, 2, 3)$, let $\mathscr{L}\colon \mathbb{R}^3 \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ that

$$\mathscr{L}(\theta) = 2(\theta_1)^2 + (\theta_2 + 1)^2 + (\theta_3 - 1)^2, \tag{6.2}$$

and let Θ be the GD process for the objective function $\mathscr{L}$ with learning rates $\mathbb{N} \ni n \mapsto \frac{1}{2^n}$, and initial value ξ (cf. Definition 6.1.1). Specify Θ_1, Θ_2, and Θ_3 explicitly and prove that your results are correct!

Exercise 6.1.2. Let $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$ satisfy $\xi = (\xi_1, \xi_2, \xi_3) = (3, 4, 5)$, let $\mathscr{L}\colon \mathbb{R}^3 \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_3) \in \mathbb{R}^3$ that

$$\mathscr{L}(\theta) = (\theta_1)^2 + (\theta_2 - 1)^2 + 2\,(\theta_3 + 1)^2,$$

and let Θ be the GD process for the objective function $\mathscr{L}$ with learning rates $\mathbb{N} \ni n \mapsto 1/3 \in [0,\infty)$ and initial value ξ (cf. Definition 6.1.1). Specify Θ_1, Θ_2, and Θ_3 explicitly and prove that your results are correct.

6.1.1 GD optimization in the training of ANNs

In the next example we apply the GD optimization method in the context of the training of fully-connected feedforward ANNs in the vectorized description (see Section 1.1) with the loss function being the mean squared error loss function in Definition 5.4.2 (see Section 5.4.2).

Example 6.1.3. *Let $d, h, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \ldots, l_h \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + \left[\sum_{k=2}^{h} l_k(l_{k-1}+1)\right] + l_h + 1$, let $a\colon \mathbb{R} \to \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $x_1, x_2, \ldots, x_M \in \mathbb{R}^d$, $y_1, y_2, \ldots, y_M \in \mathbb{R}$,*

let $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$\mathscr{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^{M} \left| \left(\mathcal{N}^{\theta,d}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\ldots,\mathfrak{M}_{a,l_h},\mathrm{id}_{\mathbb{R}}}\right)(x_m) - y_m \right|^2\right], \tag{6.3}$$

let $\xi \in \mathbb{R}^d$, *let* $(\gamma_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$, *and let* $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \textit{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla\mathscr{L})(\Theta_{n-1}) \tag{6.4}$$

(cf. Definitions 1.1.3 and 1.2.1 and Corollary 5.3.6). Then Θ *is the GD process for the objective function* $\mathscr{L}$ *with learning rates* $(\gamma_n)_{n\in\mathbb{N}}$ *and initial value* ξ.

Proof for Example 6.1.3. Note that (6.1) and (6.4) demonstrate that Θ is the GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ. The proof for Example 6.1.3 is thus complete. □

6.1.2 Euler discretizations for GF ODEs

Theorem 6.1.4 (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*

$$f(b) = \left[\sum_{n=0}^{N-1} \frac{f^{(n)}(a)(b-a)^n}{n!}\right] + \int_0^1 \frac{f^{(N)}(a+r(b-a))(b-a)^N(1-r)^{N-1}}{(N-1)!}\,\mathrm{d}r. \tag{6.5}$$

Proof of Theorem 6.1.4. Observe that the fundamental theorem of calculus assures that for all $g \in C^1([0,1],\mathbb{R})$ it holds that

$$g(1) = g(0) + \int_0^1 g'(r)\,\mathrm{d}r = g(0) + \int_0^1 \frac{g'(r)(1-r)^0}{0!}\,\mathrm{d}r. \tag{6.6}$$

Furthermore, note that integration by parts ensures that for all $n \in \mathbb{N}$, $g \in C^{n+1}([0,1],\mathbb{R})$ it holds that

$$\begin{aligned}\int_0^1 \frac{g^{(n)}(r)(1-r)^{n-1}}{(n-1)!}\,\mathrm{d}r &= -\left[\frac{g^{(n)}(r)(1-r)^n}{n!}\right]_{r=0}^{r=1} + \int_0^1 \frac{g^{(n+1)}(r)(1-r)^n}{n!}\,\mathrm{d}r \\ &= \frac{g^{(n)}(0)}{n!} + \int_0^1 \frac{g^{(n+1)}(r)(1-r)^n}{n!}\,\mathrm{d}r.\end{aligned} \tag{6.7}$$

Combining this with (6.6) and induction shows that for all $g \in C^N([0,1],\mathbb{R})$ it holds that

$$g(1) = \left[\sum_{n=0}^{N-1} \frac{g^{(n)}(0)}{n!}\right] + \int_0^1 \frac{g^{(N)}(r)(1-r)^{N-1}}{(N-1)!}\,\mathrm{d}r. \tag{6.8}$$

This establishes (6.5). The proof of Theorem 6.1.4 is thus complete. □

Lemma 6.1.5 (Local error of the Euler method). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathscr{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $x, y \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 + \int_0^t \mathscr{G}(\Theta_s)\,\mathrm{d}s, \qquad \theta = \Theta_T + \gamma \mathscr{G}(\Theta_T), \tag{6.9}$$

$$\|\mathscr{G}(x)\|_2 \leq c, \qquad \text{and} \qquad \|\mathscr{G}'(x)y\|_2 \leq c\|y\|_2 \tag{6.10}$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2\gamma^2. \tag{6.11}$$

Proof of Lemma 6.1.5. Note that the fundamental theorem of calculus, the hypothesis that $\mathscr{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$, and (6.9) establish that for all $t \in (0, \infty)$ it holds that $\Theta \in C^1([0, \infty), \mathbb{R}^{\mathfrak{d}})$ and

$$\dot{\Theta}_t = \mathscr{G}(\Theta_t). \tag{6.12}$$

Combining this with the hypothesis that $\mathscr{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$ 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

$$\ddot{\Theta}_t = \mathscr{G}'(\Theta_t)\dot{\Theta}_t = \mathscr{G}'(\Theta_t)\mathscr{G}(\Theta_t). \tag{6.13}$$

Theorem 6.1.4 and (6.12) therefore imply that

$$\begin{aligned} \Theta_{T+\gamma} &= \Theta_T + \gamma\dot{\Theta}_T + \int_0^1 (1-r)\gamma^2\ddot{\Theta}_{T+r\gamma}\,\mathrm{d}r \\ &= \Theta_T + \gamma\mathscr{G}(\Theta_T) + \gamma^2 \int_0^1 (1-r)\mathscr{G}'(\Theta_{T+r\gamma})\mathscr{G}(\Theta_{T+r\gamma})\,\mathrm{d}r. \end{aligned} \tag{6.14}$$

This and (6.9) demonstrate that

$$\begin{aligned} &\|\Theta_{T+\gamma} - \theta\|_2 \\ &= \left\| \Theta_T + \gamma\mathscr{G}(\Theta_T) + \gamma^2 \int_0^1 (1-r)\mathscr{G}'(\Theta_{T+r\gamma})\mathscr{G}(\Theta_{T+r\gamma})\,\mathrm{d}r - (\Theta_T + \gamma\mathscr{G}(\Theta_T)) \right\|_2 \\ &\leq \gamma^2 \int_0^1 (1-r)\|\mathscr{G}'(\Theta_{T+r\gamma})\mathscr{G}(\Theta_{T+r\gamma})\|_2\,\mathrm{d}r \\ &\leq c^2\gamma^2 \int_0^1 r\,\mathrm{d}r = \frac{c^2\gamma^2}{2} \leq c^2\gamma^2. \end{aligned} \tag{6.15}$$

The proof of Lemma 6.1.5 is thus complete. □

Corollary 6.1.6 (Local error of the Euler method for GF ODEs)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0,\infty)$, $\mathscr{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $x, y \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0,\infty)$ *that*

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathscr{L})(\Theta_s)\,\mathrm{d}s, \qquad \theta = \Theta_T - \gamma(\nabla \mathscr{L})(\Theta_T), \tag{6.16}$$

$$\|(\nabla \mathscr{L})(x)\|_2 \leq c, \qquad \textit{and} \qquad \|(\operatorname{Hess} \mathscr{L})(x)y\|_2 \leq c\|y\|_2 \tag{6.17}$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2\gamma^2. \tag{6.18}$$

Proof of Corollary 6.1.6. Throughout this proof, let $\mathscr{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{G}(\theta) = -(\nabla \mathscr{L})(\theta). \tag{6.19}$$

Note that the fact that for all $t \in [0,\infty)$ it holds that $\Theta_t = \Theta_0 + \int_0^t \mathscr{G}(\Theta_s)\,\mathrm{d}s$, the fact that $\theta = \Theta_T + \gamma\mathscr{G}(\Theta_T)$, the fact that for all $x \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathscr{G}(x)\|_2 \leq c$, the fact that for all $x, y \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathscr{G}'(x)y\|_2 \leq c\|y\|_2$, and Lemma 6.1.5 prove that $\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2\gamma^2$. The proof of Corollary 6.1.6 is thus complete. □

6.1.3 Lyapunov-type stability for GD optimization

Corollary 5.8.3 in Section 5.8.2 and Corollary 5.8.5 in Section 5.8.3 in Chapter 5 above, in particular, illustrate how Lyapunov-type functions can be employed to establish convergence properties for GFs. Roughly speaking, the next two results, Proposition 6.1.7 and Corollary 6.1.8 below, are the time-discrete analogons of Corollary 5.8.3 and Corollary 5.8.5, respectively.

Proposition 6.1.7 (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* $V\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$, $\Phi\colon \mathbb{R}^{\mathfrak{d}} \times [0,\infty) \to \mathbb{R}^{\mathfrak{d}}$, *and* $\varepsilon\colon [0,c] \to [0,\infty)$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0,c]$ *that*

$$V(\Phi(\theta,t)) \leq \varepsilon(t)V(\theta), \tag{6.20}$$

and let $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \textit{and} \qquad \Theta_n = \Phi(\Theta_{n-1}, \gamma_n). \tag{6.21}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$V(\Theta_n) \leq \left[\prod_{k=1}^{n} \varepsilon(\gamma_k)\right] V(\xi). \tag{6.22}$$

Proof of Proposition 6.1.7. We prove (6.22) 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 (6.22) in the base case $n = 0$. For the induction step observe that (6.21) and (6.20) ensure that for all $n \in \mathbb{N}_0$ with $V(\Theta_n) \leq (\prod_{k=1}^{n} \varepsilon(\gamma_k))V(\xi)$ it holds that

$$\begin{aligned} V(\Theta_{n+1}) &= V(\Phi(\Theta_n, \gamma_{n+1})) \leq \varepsilon(\gamma_{n+1}) V(\Theta_n) \\ &\leq \varepsilon(\gamma_{n+1}) \left(\left[\prod_{k=1}^{n} \varepsilon(\gamma_k)\right] V(\xi) \right) = \left[\prod_{k=1}^{n+1} \varepsilon(\gamma_k)\right] V(\xi). \end{aligned} \tag{6.23}$$

Induction thus establishes (6.22). The proof of Proposition 6.1.7 is thus complete. □

Corollary 6.1.8 (On quadratic Lyapunov-type functions for the GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta, \xi \in \mathbb{R}^{\mathfrak{d}}$, $c \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, c]$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\|\!|\cdot|\!\| \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty)$ be a norm, let $\varepsilon \colon [0, c] \to [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, c]$ that*

$$\|\!|\theta - t(\nabla \mathscr{L})(\theta) - \vartheta|\!\|^2 \leq \varepsilon(t) \|\!|\theta - \vartheta|\!\|^2, \tag{6.24}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathscr{L})(\Theta_{n-1}). \tag{6.25}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\|\!|\Theta_n - \vartheta|\!\| \leq \left[\prod_{k=1}^{n} [\varepsilon(\gamma_k)]^{1/2}\right] \|\!|\xi - \vartheta|\!\|. \tag{6.26}$$

Proof of Corollary 6.1.8. Throughout this proof, let $V \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ and $\Phi \colon \mathbb{R}^{\mathfrak{d}} \times [0, \infty) \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$ that

$$V(\theta) = \|\!|\theta - \vartheta|\!\|^2 \qquad \text{and} \qquad \Phi(\theta, t) = \theta - t(\nabla \mathscr{L})(\theta). \tag{6.27}$$

Observe that Proposition 6.1.7 (applied with $V \curvearrowleft V$, $\Phi \curvearrowleft \Phi$ in the notation of Proposition 6.1.7) and (6.27) imply that for all $n \in \mathbb{N}_0$ it holds that

$$\|\!|\Theta_n - \vartheta|\!\|^2 = V(\Theta_n) \leq \left[\prod_{k=1}^{n} \varepsilon(\gamma_k)\right] V(\xi) = \left[\prod_{k=1}^{n} \varepsilon(\gamma_k)\right] \|\!|\xi - \vartheta|\!\|^2. \tag{6.28}$$

This establishes (6.26). The proof of Corollary 6.1.8 is thus complete. □

Corollary 6.1.8, in particular, illustrates that the one-step Lyapunov stability assumption in (6.24) may provide us suitable estimates for the approximation errors associated to the GD optimization method; see (6.26) above. The next result, Lemma 6.1.9 below, now provides us sufficient conditions which ensure that the one-step Lyapunov stability condition in (6.24) is satisfied so that we are in the position to apply Corollary 6.1.8 above to obtain estimates for the approximation errors associated to the GD optimization method. Lemma 6.1.9 employs the growth condition and the coercivity-type condition in (5.249) in Corollary 5.9.5 above. Results similar to Lemma 6.1.9 can, for instance, be found in [108, Remark 2.1] and [229, Lemma 2.1]. We will employ the statement of Lemma 6.1.9 in our error analysis for the GD optimization method in Section 6.1.4 below.

Lemma 6.1.9 (Sufficient conditions for a one-step Lyapunov-type stability condition)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle \colon \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a scalar product, let $\|\!|\cdot|\!\| \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that $\|\!|v|\!\| = \sqrt{\langle\langle v, v \rangle\rangle}$, and let $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|\!|w - \vartheta|\!\| \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle\rangle \geq c \|\!|\theta - \vartheta|\!\|^2 \qquad \text{and} \qquad \|\!|(\nabla \mathscr{L})(\theta)|\!\| \leq L \|\!|\theta - \vartheta|\!\|. \tag{6.29}$$

Then

(i) it holds that $c \leq L$,

(ii) it holds for all $\theta \in \mathbb{B}$, $\gamma \in [0, \infty)$ that

$$\|\!|\theta - \gamma(\nabla \mathscr{L})(\theta) - \vartheta|\!\|^2 \leq (1 - 2\gamma c + \gamma^2 L^2) \|\!|\theta - \vartheta|\!\|^2, \tag{6.30}$$

(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

$$\|\!|\theta - \gamma(\nabla \mathscr{L})(\theta) - \vartheta|\!\|^2 \leq (1 - c\gamma) \|\!|\theta - \vartheta|\!\|^2. \tag{6.31}$$

Proof of Lemma 6.1.9. First of all, note that (6.29) ensures that for all $\theta \in \mathbb{B}$, $\gamma \in [0, \infty)$ it holds that

$$\begin{aligned} 0 \leq \|\!|\theta - \gamma(\nabla \mathscr{L})(\theta) - \vartheta|\!\|^2 &= \|\!|(\theta - \vartheta) - \gamma(\nabla \mathscr{L})(\theta)|\!\|^2 \\ &= \|\!|\theta - \vartheta|\!\|^2 - 2\gamma \langle\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle\rangle + \gamma^2 \|\!|(\nabla \mathscr{L})(\theta)|\!\|^2 \\ &\leq \|\!|\theta - \vartheta|\!\|^2 - 2\gamma c \|\!|\theta - \vartheta|\!\|^2 + \gamma^2 L^2 \|\!|\theta - \vartheta|\!\|^2 \\ &= (1 - 2\gamma c + \gamma^2 L^2) \|\!|\theta - \vartheta|\!\|^2. \end{aligned} \tag{6.32}$$

This establishes item (ii). Moreover, note that the fact that $\mathbb{B} \backslash \{\vartheta\} \neq \emptyset$ and (6.32) assure that for all $\gamma \in [0, \infty)$ it holds that

$$1 - 2\gamma c + \gamma^2 L^2 \geq 0. \tag{6.33}$$

Hence, we obtain that

$$\begin{aligned} 1 - \tfrac{c^2}{L^2} = 1 - \tfrac{2c^2}{L^2} + \tfrac{c^2}{L^2} &= 1 - 2\big[\tfrac{c}{L^2}\big]c + \big[\tfrac{c^2}{L^4}\big]L^2 \\ &= 1 - 2\big[\tfrac{c}{L^2}\big]c + \big[\tfrac{c}{L^2}\big]^2 L^2 \geq 0. \end{aligned} \tag{6.34}$$

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 (6.33) ensures that for all $\gamma \in (0, \frac{2c}{L^2})$ it holds that

$$0 \leq 1 - 2\gamma c + \gamma^2 L^2 = 1 - \underbrace{\gamma}_{>0}\underbrace{(2c - \gamma L^2)}_{>0} < 1. \tag{6.35}$$

This proves item (iii). In addition, note that for all $\gamma \in [0, \frac{c}{L^2}]$ it holds that

$$1 - 2\gamma c + \gamma^2 L^2 \leq 1 - 2\gamma c + \gamma\big[\tfrac{c}{L^2}\big]L^2 = 1 - c\gamma. \tag{6.36}$$

Combining this with (6.32) establishes item (iv). The proof of Lemma 6.1.9 is thus complete. □

Exercise 6.1.3. Prove or disprove the following statement: There exist $\mathfrak{d} \in \mathbb{N}$, $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$, $r \in (0, \infty]$, $\vartheta, \theta \in \mathbb{R}^{\mathfrak{d}}$ and there exists a function $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ such that $\|\theta - \vartheta\|_2 \leq r$, $\forall\, \xi \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\} \colon \|\xi - \gamma\mathscr{G}(\xi) - \vartheta\|_2 \leq \varepsilon\|\xi - \vartheta\|_2$, and

$$\langle \theta - \vartheta, \mathscr{G}(\theta)\rangle < \min\big\{\tfrac{1-\varepsilon^2}{2\gamma}, \tfrac{\gamma}{2}\big\} \max\big\{\|\theta - \vartheta\|_2^2, \|\mathscr{G}(\theta)\|_2^2\big\}. \tag{6.37}$$

Exercise 6.1.4. Prove or disprove the following statement: For all $\mathfrak{d} \in \mathbb{N}$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and for every function $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ which satisfies $\forall\, \theta \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\} \colon \langle \theta - \vartheta, \mathscr{G}(\theta)\rangle \geq \frac{1}{2}\max\{\|\theta - \vartheta\|_2^2, \|\mathscr{G}(\theta)\|_2^2\}$ it holds that

$$\forall \theta \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\} \colon \big(\langle \theta - \vartheta, \mathscr{G}(\theta)\rangle \geq \tfrac{1}{2}\|\theta - \vartheta\|_2^2 \wedge \|\mathscr{G}(\theta)\|_2 \leq 2\|\theta - \vartheta\|_2\big). \tag{6.38}$$

Exercise 6.1.5. Prove or disprove the following statement: For all $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta, v \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $s, t \in [0, 1]$ such that $\|v\|_2 \leq r$, $s \leq t$, and $\forall\, \theta \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\} \colon \langle \theta - \vartheta, (\nabla\mathscr{L})(\theta)\rangle \geq c\|\theta - \vartheta\|_2^2$ it holds that

$$\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta + sv) \geq \tfrac{c}{2}(t^2 - s^2)\|v\|_2^2. \tag{6.39}$$

Exercise 6.1.6. Prove or disprove the following statement: For every $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and for every $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ which satisfies for all $v \in \mathbb{R}^{\mathfrak{d}}$, $s, t \in [0, 1]$ with $\|v\|_2 \leq r$ and $s \leq t$ that $\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta + sv) \geq c(t^2 - s^2)\|v\|_2^2$ it holds that

$$\forall\, \theta \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\} \colon \langle \theta - \vartheta, (\nabla\mathscr{L})(\theta)\rangle \geq 2c\|\theta - \vartheta\|_2^2. \tag{6.40}$$

Exercise 6.1.7. Let $\mathfrak{d} \in \mathbb{N}$ and for every $v \in \mathbb{R}^{\mathfrak{d}}$, $R \in [0, \infty]$ let $\mathbb{B}_R(v) = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - v\|_2 \leq R\}$. Prove or disprove the following statement: For all $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ the following two statements are equivalent:

(i) There exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{B}_r(\vartheta)$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2. \tag{6.41}$$

(ii) There exists $c \in (0, \infty)$ such that for all $v, w \in \mathbb{B}_r(\vartheta)$, $s, t \in [0, 1]$ with $s \leq t$ it holds that

$$\mathscr{L}(\vartheta + t(v - \vartheta)) - \mathscr{L}(\vartheta + s(v - \vartheta)) \geq c(t^2 - s^2)\|v - \vartheta\|_2^2. \tag{6.42}$$

Exercise 6.1.8. Let $\mathfrak{d} \in \mathbb{N}$ and for every $v \in \mathbb{R}^{\mathfrak{d}}$, $R \in [0, \infty]$ let $\mathbb{B}_R(v) = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|v - w\|_2 \leq R\}$. Prove or disprove the following statement: For all $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ the following three statements are equivalent:

(i) There exist $c, L \in (0, \infty)$ such that for all $\theta \in \mathbb{B}_r(\vartheta)$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2. \tag{6.43}$$

(ii) There exist $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$ such that for all $\theta \in \mathbb{B}_r(\vartheta)$ it holds that

$$\|\theta - \gamma(\nabla \mathscr{L})(\theta) - \vartheta\|_2 \leq \varepsilon \|\theta - \vartheta\|_2. \tag{6.44}$$

(iii) There exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{B}_r(\vartheta)$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \max\{\|\theta - \vartheta\|_2^2, \|(\nabla \mathscr{L})(\theta)\|_2^2\}. \tag{6.45}$$

6.1.4 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 6.1.10 below) that the considered GD process converges to a local minimum point of the objective function of the considered optimization problem.

6.1.4.1 Error estimates for GD optimization

Proposition 6.1.10 (Error estimates for the GD optimization method)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{2c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *satisfy for all* $\theta \in \mathbb{B}$ *that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \qquad \textit{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \tag{6.46}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \textit{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.47}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) *it holds that* $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w \in \mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(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*

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^{n} (1 - 2c\gamma_k + (\gamma_k)^2 L^2)^{1/2}\right] \|\xi - \vartheta\|_2, \tag{6.48}$$

and

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

$$0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \tfrac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \tfrac{L}{2} \left[\prod_{k=1}^{n} (1 - 2c\gamma_k + (\gamma_k)^2 L^2)\right] \|\xi - \vartheta\|_2^2. \tag{6.49}$$

Proof of Proposition 6.1.10. First, observe that (6.46) and item (ii) in Lemma 5.7.22 prove item (i). Moreover, note that (6.46), item (iii) in Lemma 6.1.9, the assumption that for all $n \in \mathbb{N}$ it holds that $\gamma_n \in [0, \frac{2c}{L^2}]$, and the fact that

$$1 - 2c\left[\tfrac{2c}{L^2}\right] + \left[\tfrac{2c}{L^2}\right]^2 L^2 = 1 - \tfrac{4c^2}{L^2} + \left[\tfrac{4c^2}{L^4}\right] L^2 = 1 - \tfrac{4c^2}{L^2} + \tfrac{4c^2}{L^2} = 1 \tag{6.50}$$

and establish item (ii). Next we claim that for all $n \in \mathbb{N}$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq (1 - 2c\gamma_n + (\gamma_n)^2 L^2)^{1/2} \|\Theta_{n-1} - \vartheta\|_2 \leq r. \tag{6.51}$$

We now prove (6.51) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ observe that (6.47), the assumption that $\Theta_0 = \xi \in \mathbb{B}$, item (ii) in Lemma 6.1.9, and item (ii) ensure that

$$\begin{aligned} \|\Theta_1 - \vartheta\|_2^2 &= \|\Theta_0 - \gamma_1 (\nabla \mathscr{L})(\Theta_0) - \vartheta\|_2^2 \\ &\leq (1 - 2c\gamma_1 + (\gamma_1)^2 L^2) \|\Theta_0 - \vartheta\|_2^2 \\ &\leq \|\Theta_0 - \vartheta\|_2^2 \leq r^2. \end{aligned} \tag{6.52}$$

This establishes (6.51) in the base case $n = 1$. For the induction step note that (6.47), item (ii) in Lemma 6.1.9, and item (ii) imply that for all $n \in \mathbb{N}$ with $\Theta_n \in \mathbb{B}$ it holds that

$$\begin{aligned} \|\Theta_{n+1} - \vartheta\|_2^2 &= \|\Theta_n - \gamma_{n+1} (\nabla \mathscr{L})(\Theta_n) - \vartheta\|_2^2 \\ &\leq \underbrace{(1 - 2c\gamma_{n+1} + (\gamma_{n+1})^2 L^2)}_{\in [0,1]} \|\Theta_n - \vartheta\|_2^2 \\ &\leq \|\Theta_n - \vartheta\|_2^2 \leq r^2. \end{aligned} \tag{6.53}$$

This demonstrates that for all $n \in \mathbb{N}$ with $\|\Theta_n - \vartheta\|_2 \leq r$ it holds that

$$\|\Theta_{n+1} - \vartheta\|_2 \leq (1 - 2c\gamma_{n+1} + (\gamma_{n+1})^2 L^2)^{1/2} \|\Theta_n - \vartheta\|_2 \leq r. \tag{6.54}$$

Induction thus proves (6.51). Next observe that (6.51) establishes item (iii). Moreover, note that induction, item (ii), and item (iii) prove item (iv). Furthermore, observe 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, (6.46), and Lemma 5.8.6 with items (i) and (iv) establishes item (v). The proof of Proposition 6.1.10 is thus complete. □

6.1.4.2 Size of the learning rates

In the next result, Corollary 6.1.11 below, we, roughly speaking, specialize Proposition 6.1.10 to the case where the learning rates $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0, \frac{2c}{L^2}]$ are a constant sequence.

Corollary 6.1.11 (Convergence of GD for constant learning rates)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\gamma \in (0, \frac{2c}{L^2})$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2, \tag{6.55}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma(\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.56}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w\in\mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(ii) it holds that $0 \leq 1 - 2c\gamma + \gamma^2 L^2 < 1$,

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

$$\|\Theta_n - \vartheta\|_2 \leq \left[1 - 2c\gamma + \gamma^2 L^2\right]^{n/2} \|\xi - \vartheta\|_2, \tag{6.57}$$

and

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

$$0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \tfrac{L}{2}\|\Theta_n - \vartheta\|_2^2 \leq \tfrac{L}{2}\left[1 - 2c\gamma + \gamma^2 L^2\right]^n \|\xi - \vartheta\|_2^2. \tag{6.58}$$

Proof of Corollary 6.1.11. Observe that item (iii) in Lemma 6.1.9 proves item (ii). In addition, note that Proposition 6.1.10 establishes items (i), (iii), and (iv). The proof of Corollary 6.1.11 is thus complete. □

Corollary 6.1.11 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 6.1.12 below, demonstrates that the condition that the learning rates are strictly smaller than $\frac{2c}{L^2}$ in Corollary 6.1.11 can, in general, not be relaxed.

Theorem 6.1.12 (Sharp bounds on the learning rate for the convergence of GD)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\alpha \in (0, \infty)$, $\gamma \in \mathbb{R}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\xi \in \mathbb{R}^{\mathfrak{d}} \backslash \{\vartheta\}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \tfrac{\alpha}{2} \|\theta - \vartheta\|_2^2, \tag{6.59}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma (\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.60}$$

(cf. Definition 3.3.4). Then

(i) it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle = \alpha \|\theta - \vartheta\|_2^2$,

(ii) it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\|(\nabla \mathscr{L})(\theta)\|_2 = \alpha \|\theta - \vartheta\|_2$,

(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

$$\liminf_{n \to \infty} \|\Theta_n - \vartheta\|_2 = \limsup_{n \to \infty} \|\Theta_n - \vartheta\|_2 = \begin{cases} 0 & : \gamma \in (0, 2/\alpha) \\ \|\xi - \vartheta\|_2 & : \gamma \in \{0, 2/\alpha\} \\ \infty & : \gamma \in \mathbb{R} \backslash [0, 2/\alpha] \end{cases} \tag{6.61}$$

(cf. Definition 1.4.7).

Proof of Theorem 6.1.12. First of all, note that Lemma 5.8.4 ensures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\mathscr{L} \in C^\infty(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and

$$(\nabla \mathscr{L})(\theta) = \tfrac{\alpha}{2}(2(\theta - \vartheta)) = \alpha(\theta - \vartheta). \tag{6.62}$$

This proves item (ii). Moreover, observe that (6.62) assures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle = \langle \theta - \vartheta, \alpha(\theta - \vartheta) \rangle = \alpha \|\theta - \vartheta\|_2^2 \tag{6.63}$$

(cf. Definition 1.4.7). This establishes item (i). Observe that (6.60) and (6.62) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Theta_n - \vartheta &= \Theta_{n-1} - \gamma (\nabla \mathscr{L})(\Theta_{n-1}) - \vartheta \\ &= \Theta_{n-1} - \gamma\alpha(\Theta_{n-1} - \vartheta) - \vartheta \\ &= (1 - \gamma\alpha)(\Theta_{n-1} - \vartheta). \end{aligned} \tag{6.64}$$

The assumption that $\Theta_0 = \xi$ and induction hence prove that for all $n \in \mathbb{N}_0$ it holds that

$$\Theta_n - \vartheta = (1 - \gamma\alpha)^n(\Theta_0 - \vartheta) = (1 - \gamma\alpha)^n(\xi - \vartheta). \tag{6.65}$$

Therefore, we obtain for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 = |1 - \gamma\alpha|^n \|\xi - \vartheta\|_2. \tag{6.66}$$

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 $t \in \mathbb{R}\backslash[0, 2/\alpha]$ it holds that $|1 - t\alpha| \in (1, \infty)$, and the fact that $\|\xi - \vartheta\|_2 > 0$ establishes item (iv). The proof of Theorem 6.1.12 is thus complete. □

Exercise 6.1.9. Let $\mathscr{L} \colon \mathbb{R} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathscr{L}(\theta) = 2\theta^2 \tag{6.67}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0 = 1$ and

$$\Theta_n = \Theta_{n-1} - n^{-2}(\nabla\mathscr{L})(\Theta_{n-1}). \tag{6.68}$$

Prove or disprove the following statement: It holds that

$$\limsup_{n\to\infty} |\Theta_n| = 0. \tag{6.69}$$

Exercise 6.1.10. Let $\mathscr{L} \colon \mathbb{R} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathscr{L}(\theta) = 4\theta^2 \tag{6.70}$$

and for every $r \in (1, \infty)$ let $\Theta^{(r)} \colon \mathbb{N}_0 \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0^{(r)} = 1$ and

$$\Theta_n^{(r)} = \Theta_{n-1}^{(r)} - n^{-r}(\nabla\mathscr{L})(\Theta_{n-1}^{(r)}). \tag{6.71}$$

Prove or disprove the following statement: It holds for all $r \in (1, \infty)$ that

$$\liminf_{n\to\infty} |\Theta_n^{(r)}| > 0. \tag{6.72}$$

Exercise 6.1.11. Let $\mathscr{L} \colon \mathbb{R} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathscr{L}(\theta) = 5\theta^2 \tag{6.73}$$

and for every $r \in (1, \infty)$ let $\Theta^{(r)} = (\Theta_n^{(r)})_{n\in\mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0^{(r)} = 1$ and

$$\Theta_n^{(r)} = \Theta_{n-1}^{(r)} - n^{-r}(\nabla\mathscr{L})(\Theta_{n-1}^{(r)}). \tag{6.74}$$

Prove or disprove the following statement: It holds for all $r \in (1, \infty)$ that

$$\liminf_{n\to\infty} |\Theta_n^{(r)}| > 0. \tag{6.75}$$

6.1.4.3 Convergence rates

The next result, Corollary 6.1.13 below, establishes a convergence rate for the GD optimization method in the case of possibly non-constant learning rates. We prove Corollary 6.1.13 through an application of Proposition 6.1.10 above.

Corollary 6.1.13 (Qualitative convergence of GD). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$, $c, L \in (0,\infty)$, $\xi, \vartheta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta)\rangle \geq c\|\theta - \vartheta\|_2^2, \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2, \tag{6.76}$$

$$\text{and} \qquad 0 < \liminf_{n\to\infty} \gamma_n \leq \limsup_{n\to\infty} \gamma_n < \tfrac{2c}{L^2}, \tag{6.77}$$

and let $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.78}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{R}^{\mathfrak{d}}\colon \mathscr{L}(\theta) = \inf_{w\in\mathbb{R}^{\mathfrak{d}}} \mathscr{L}(w)\} = \{\vartheta\}$,

(ii) there exist $\epsilon \in (0,1)$, $C \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \epsilon^n C, \tag{6.79}$$

and

(iii) there exist $\epsilon \in (0,1)$, $C \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \epsilon^n C. \tag{6.80}$$

Proof of Corollary 6.1.13. Throughout this proof, let $\alpha, \beta \in \mathbb{R}$ satisfy

$$0 < \alpha < \liminf_{n\to\infty} \gamma_n \leq \limsup_{n\to\infty} \gamma_n < \beta < \tfrac{2c}{L^2} \tag{6.81}$$

(cf. (6.77)), let $m \in \mathbb{N}$ satisfy for all $n \in \mathbb{N}$ that $\gamma_{m+n} \in [\alpha, \beta]$, and let $h\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $t \in \mathbb{R}$ that

$$h(t) = 1 - 2ct + t^2L^2. \tag{6.82}$$

Observe that (6.76) and item (ii) in Lemma 5.7.22 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

$$h'(t) \leq -2c + 2\left[\tfrac{c}{L^2}\right]L^2 = 0. \tag{6.83}$$

The fundamental theorem of calculus hence assures that for all $t \in [\alpha, \beta] \cap (-\infty, \frac{c}{L^2}]$ it holds that

$$h(t) = h(\alpha) + \int_\alpha^t h'(s)\,\mathrm{d}s \leq h(\alpha) + \int_\alpha^t 0\,\mathrm{d}s = h(\alpha) \leq \max\{h(\alpha), h(\beta)\}. \tag{6.84}$$

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

$$h'(t) \geq h'(\tfrac{c}{L^2}) = -2c + 2\left[\tfrac{c}{L^2}\right]L^2 = 0. \tag{6.85}$$

The fundamental theorem of calculus hence ensures that for all $t \in [\alpha, \beta] \cap [\frac{c}{L^2}, \infty)$ it holds that

$$\max\{h(\alpha), h(\beta)\} \geq h(\beta) = h(t) + \int_t^\beta h'(s)\,\mathrm{d}s \geq h(t) + \int_t^\beta 0\,\mathrm{d}s = h(t). \tag{6.86}$$

Combining this and (6.84) establishes that for all $t \in [\alpha, \beta]$ it holds that

$$h(t) \leq \max\{h(\alpha), h(\beta)\}. \tag{6.87}$$

Moreover, observe that the fact that $\alpha, \beta \in (0, \frac{2c}{L^2})$ and item (iii) in Lemma 6.1.9 ensure that

$$\{h(\alpha), h(\beta)\} \subseteq [0, 1). \tag{6.88}$$

Hence, we obtain that

$$\max\{h(\alpha), h(\beta)\} \in [0, 1). \tag{6.89}$$

This implies that there exists $\varepsilon \in \mathbb{R}$ such that

$$0 \leq \max\{h(\alpha), h(\beta)\} < \varepsilon < 1. \tag{6.90}$$

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 6.1.10 (applied with $\mathfrak{d} \curvearrowleft \mathfrak{d}$, $c \curvearrowleft c$, $L \curvearrowleft L$, $r \curvearrowleft \infty$, $(\gamma_n)_{n\in\mathbb{N}} \curvearrowleft (\gamma_{m+n})_{n\in\mathbb{N}}$, $\vartheta \curvearrowleft \vartheta$, $\xi \curvearrowleft \Theta_m$, $\mathscr{L} \curvearrowleft \mathscr{L}$ in the notation of Proposition 6.1.10), (6.76), (6.78), and (6.87) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
\|\Theta_{m+n} - \vartheta\|_2 &\leq \left[\prod_{k=1}^{n} (1 - 2c\gamma_{m+k} + (\gamma_{m+k})^2 L^2)^{1/2}\right] \|\Theta_m - \vartheta\|_2 \\
&= \left[\prod_{k=1}^{n} (h(\gamma_{m+k}))^{1/2}\right] \|\Theta_m - \vartheta\|_2 \\
&\leq (\max\{h(\alpha), h(\beta)\})^{n/2} \|\Theta_m - \vartheta\|_2 \\
&\leq \varepsilon^{n/2} \|\Theta_m - \vartheta\|_2.
\end{aligned} \tag{6.91}$$

This shows that for all $n \in \mathbb{N}$ with $n > m$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \varepsilon^{(n-m)/2} \|\Theta_m - \vartheta\|_2. \tag{6.92}$$

The fact that for all $n \in \mathbb{N}_0$ with $n \leq m$ it holds that

$$\|\Theta_n - \vartheta\|_2 = \left[\frac{\|\Theta_n - \vartheta\|_2}{\varepsilon^{n/2}}\right] \varepsilon^{n/2} \leq \left[\max\left\{\frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} \colon k \in \{0, 1, \ldots, m\}\right\}\right] \varepsilon^{n/2} \tag{6.93}$$

hence assures that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
\|\Theta_n - \vartheta\|_2 &\leq \max\left\{\left[\max\left\{\frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} \colon k \in \{0, 1, \ldots, m\}\right\}\right] \varepsilon^{n/2}, \varepsilon^{(n-m)/2} \|\Theta_m - \vartheta\|_2\right\} \\
&= (\varepsilon^{1/2})^n \left[\max\left\{\max\left\{\frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} \colon k \in \{0, 1, \ldots, m\}\right\}, \varepsilon^{-m/2} \|\Theta_m - \vartheta\|_2\right\}\right] \\
&= (\varepsilon^{1/2})^n \left[\max\left\{\frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} \colon k \in \{0, 1, \ldots, m\}\right\}\right].
\end{aligned} \tag{6.94}$$

This proves item (ii). In addition, note that Lemma 5.8.6, item (i), and (6.94) assure that for all $n \in \mathbb{N}_0$ it holds that

$$0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \tfrac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{\varepsilon^n L}{2} \left[\max\left\{\frac{\|\Theta_k - \vartheta\|_2^2}{\varepsilon^k} \colon k \in \{0, 1, \ldots, m\}\right\}\right]. \tag{6.95}$$

This establishes item (iii). The proof of Corollary 6.1.13 is thus complete. □

6.1.4.4 Error estimates in the case of small learning rates

The inequality in (6.48) in item (iv) in Proposition 6.1.10 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 6.1.10 satisfy that for all $n \in \mathbb{N}$ it holds that $\gamma_n \leq \frac{2c}{L^2}$. The error estimate in (6.48) 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 6.1.14 below. We prove Corollary 6.1.14 through an application of Proposition 6.1.10 above.

Corollary 6.1.14 (Error estimates in the case of small learning rates)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \tag{6.96}$$

and let $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.97}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{B}\colon \mathscr{L}(\theta) = \inf_{w\in\mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(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*

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^{n} (1 - c\gamma_k)^{1/2}\right] \|\xi - \vartheta\|_2 \leq \exp\!\left(-\tfrac{c}{2}\left[\textstyle\sum_{k=1}^{n} \gamma_k\right]\right) \|\xi - \vartheta\|_2, \tag{6.98}$$

and

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

$$0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \tfrac{L}{2}\left[\prod_{k=1}^{n} (1 - c\gamma_k)\right] \|\xi - \vartheta\|_2^2 \leq \tfrac{L}{2} \exp\!\left(-c\left[\textstyle\sum_{k=1}^{n} \gamma_k\right]\right) \|\xi - \vartheta\|_2^2. \tag{6.99}$$

Proof of Corollary 6.1.14. Note that item (ii) in Proposition 6.1.10 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

$$0 \leq 1 - 2c\gamma_n + (\gamma_n)^2 L^2 \leq 1 - 2c\gamma_n + \gamma_n \left[\frac{c}{L^2}\right] L^2 = 1 - 2c\gamma_n + \gamma_n c = 1 - c\gamma_n \leq 1. \tag{6.100}$$

This proves item (ii). Moreover, note that (6.100) and Proposition 6.1.10 establish items (i), (iii), and (iv). The proof of Corollary 6.1.14 is thus complete. □

In the next result, Corollary 6.1.15 below, we, roughly speaking, specialize Corollary 6.1.14 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 6.1.15 (Error estimates in the case of small and constant learning rates)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\gamma \in (0, \frac{c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}}\colon \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *satisfy for all* $\theta \in \mathbb{B}$ *that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta)\rangle \geq c\|\theta - \vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2, \tag{6.101}$$

and let $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma(\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.102}$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w \in \mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(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* $0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \frac{L}{2} (1 - c\gamma)^n \|\xi - \vartheta\|_2^2$.

Proof of Corollary 6.1.15. Corollary 6.1.15 is an immediate consequence of Corollary 6.1.14. The proof of Corollary 6.1.15 is thus complete. □

6.1.4.5 On the spectrum of the Hessian of the objective function at a local minimum point

A crucial ingredient in our error analysis for the GD optimization method in Sections 6.1.4.1, 6.1.4.2, 6.1.4.3, and 6.1.4.4 above is to employ the growth and the coercivity-type hypotheses, for example, in (6.46) in Proposition 6.1.10 above. In this subsection we disclose in Proposition 6.1.17 below suitable conditions on the Hessians of the objective function of the considered optimization problem which are sufficient to ensure that (6.46) is satisfied so that we are in the position to apply the error analysis in Sections 6.1.4.1, 6.1.4.2, 6.1.4.3, and 6.1.4.4 above (cf. Corollary 6.1.18 below). Our proof of Proposition 6.1.17 employs the following classical result (see Lemma 6.1.16 below) for symmetric matrices with real entries.

Lemma 6.1.16 (Properties of the spectrum of real symmetric matrices)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, *let* $A \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ *be a symmetric matrix, and let*

$$\mathcal{S} = \{\lambda \in \mathbb{C} \colon (\exists\, v \in \mathbb{C}^{\mathfrak{d}} \backslash \{0\} \colon Av = \lambda v)\}. \tag{6.103}$$

Then

(i) it holds that $\mathcal{S} = \{\lambda \in \mathbb{R} \colon (\exists\, v \in \mathbb{R}^{\mathfrak{d}} \backslash \{0\} \colon Av = \lambda v)\} \subseteq \mathbb{R}$,

(ii) it holds that

$$\sup_{v \in \mathbb{R}^{\mathfrak{d}} \backslash \{0\}} \left[\frac{\|Av\|_2}{\|v\|_2} \right] = \max_{\lambda \in \mathcal{S}} |\lambda|, \tag{6.104}$$

and

(iii) it holds for all $v \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$\min(\mathcal{S}) \|v\|_2^2 \leq \langle v, Av \rangle \leq \max(\mathcal{S}) \|v\|_2^2 \tag{6.105}$$

(cf. Definitions 1.4.7 and 3.3.4).

Proof of Lemma 6.1.16. Throughout this proof, let $e_1, e_2, \ldots, e_{\mathfrak{d}} \in \mathbb{R}^{\mathfrak{d}}$ be the vectors given by

$$e_1 = (1, 0, \ldots, 0), \qquad e_2 = (0, 1, 0, \ldots, 0), \qquad \ldots, \qquad e_{\mathfrak{d}} = (0, \ldots, 0, 1). \tag{6.106}$$

Observe that the spectral theorem for symmetric matrices (see, for instance, Petersen [345, Theorem 4.3.4]) proves that there exist $(\mathfrak{d} \times \mathfrak{d})$-matrices $\Lambda = (\Lambda_{i,j})_{(i,j)\in\{1,2,\ldots,\mathfrak{d}\}^2}$, $O = (O_{i,j})_{(i,j)\in\{1,2,\ldots,\mathfrak{d}\}^2} \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}}$ such that $\mathcal{S} = \{\Lambda_{1,1}, \Lambda_{2,2}, \ldots, \Lambda_{\mathfrak{d},\mathfrak{d}}\}$, $O^*O = OO^* = \mathrm{I}_{\mathfrak{d}}$, $A = O\Lambda O^*$, and

$$\Lambda = \begin{pmatrix} \Lambda_{1,1} & & 0 \\ & \ddots & \\ 0 & & \Lambda_{\mathfrak{d},\mathfrak{d}} \end{pmatrix} \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}} \tag{6.107}$$

(cf. Definition 1.5.5). Hence, we obtain that $\mathcal{S} \subseteq \mathbb{R}$. Next note that the assumption that $\mathcal{S} = \{\lambda \in \mathbb{C} \colon (\exists\, v \in \mathbb{C}^{\mathfrak{d}}\backslash\{0\} \colon Av = \lambda v)\}$ ensures that for every $\lambda \in \mathcal{S}$ there exists $v \in \mathbb{C}^{\mathfrak{d}}\backslash\{0\}$ such that

$$A\mathfrak{Re}(v) + \mathbf{i}A\mathfrak{Im}(v) = Av = \lambda v = \lambda\mathfrak{Re}(v) + \mathbf{i}\lambda\mathfrak{Im}(v). \tag{6.108}$$

The fact that $\mathcal{S} \subseteq \mathbb{R}$ therefore demonstrates that for every $\lambda \in \mathcal{S}$ there exists $v \in \mathbb{R}^{\mathfrak{d}}\backslash\{0\}$ such that $Av = \lambda v$. This and the fact that $\mathcal{S} \subseteq \mathbb{R}$ ensure that $\mathcal{S} \subseteq \{\lambda \in \mathbb{R} \colon (\exists\, v \in \mathbb{R}^{\mathfrak{d}}\backslash\{0\} \colon Av = \lambda v)\}$. Combining this and the fact that $\{\lambda \in \mathbb{R} \colon (\exists\, v \in \mathbb{R}^{\mathfrak{d}}\backslash\{0\} \colon Av = \lambda v)\} \subseteq \mathcal{S}$ proves item (i). Furthermore, note that (6.107) assures that for all $v = (v_1, v_2, \ldots, v_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \|\Lambda v\|_2 &= \left[\sum_{i=1}^{\mathfrak{d}} |\Lambda_{i,i} v_i|^2\right]^{1/2} \leq \left[\sum_{i=1}^{\mathfrak{d}} \max\{|\Lambda_{1,1}|^2, \ldots, |\Lambda_{\mathfrak{d},\mathfrak{d}}|^2\} |v_i|^2\right]^{1/2} \\ &= \left[\max\{|\Lambda_{1,1}|, \ldots, |\Lambda_{\mathfrak{d},\mathfrak{d}}|\}^2 \|v\|_2^2\right]^{1/2} \\ &= \max\{|\Lambda_{1,1}|, \ldots, |\Lambda_{\mathfrak{d},\mathfrak{d}}|\} \|v\|_2 \\ &= \left(\max_{\lambda\in\mathcal{S}} |\lambda|\right) \|v\|_2 \end{aligned} \tag{6.109}$$

(cf. Definition 3.3.4). The fact that O is an orthogonal matrix and the fact that $A = O\Lambda O^*$ therefore imply that for all $v \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \|Av\|_2 &= \|O\Lambda O^* v\|_2 = \|\Lambda O^* v\|_2 \\ &\leq \left(\max_{\lambda\in\mathcal{S}} |\lambda|\right) \|O^* v\|_2 \\ &= \left(\max_{\lambda\in\mathcal{S}} |\lambda|\right) \|v\|_2. \end{aligned} \tag{6.110}$$

This implies that

$$\sup_{v\in\mathbb{R}^{\mathfrak{d}}\setminus\{0\}}\left[\frac{\|Av\|_2}{\|v\|_2}\right]\leq\sup_{v\in\mathbb{R}^{\mathfrak{d}}\setminus\{0\}}\left[\frac{(\max_{\lambda\in\mathcal{S}}|\lambda|)\|v\|_2}{\|v\|_2}\right]=\max_{\lambda\in\mathcal{S}}|\lambda|. \tag{6.111}$$

In addition, note that the fact that $\mathcal{S}=\{\Lambda_{1,1},\Lambda_{2,2}\ldots,\Lambda_{\mathfrak{d},\mathfrak{d}}\}$ ensures that there exists $j\in\{1,2,\ldots,\mathfrak{d}\}$ such that

$$|\Lambda_{j,j}|=\max_{\lambda\in\mathcal{S}}|\lambda|. \tag{6.112}$$

Next observe that the fact that $A=O\Lambda O^*$, the fact that O is an orthogonal matrix, and (6.112) imply that

$$\begin{aligned}\sup_{v\in\mathbb{R}^{\mathfrak{d}}\setminus\{0\}}\left[\frac{\|Av\|_2}{\|v\|_2}\right]&\geq\frac{\|AOe_j\|_2}{\|Oe_j\|_2}=\|O\Lambda O^*Oe_j\|_2=\|O\Lambda e_j\|_2\\&=\|\Lambda e_j\|_2=\|\Lambda_{j,j}e_j\|_2=|\Lambda_{j,j}|=\max_{\lambda\in\mathcal{S}}|\lambda|.\end{aligned} \tag{6.113}$$

Combining this and (6.111) establishes item (ii). It thus remains to prove item (iii). For this note that (6.107) ensures that for all $v=(v_1,v_2,\ldots,v_{\mathfrak{d}})\in\mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}\langle v,\Lambda v\rangle&=\sum_{i=1}^{\mathfrak{d}}\Lambda_{i,i}|v_i|^2\leq\sum_{i=1}^{\mathfrak{d}}\max\{\Lambda_{1,1},\ldots,\Lambda_{\mathfrak{d},\mathfrak{d}}\}|v_i|^2\\&=\max\{\Lambda_{1,1},\ldots,\Lambda_{\mathfrak{d},\mathfrak{d}}\}\|v\|_2^2=\max(\mathcal{S})\|v\|_2^2\end{aligned} \tag{6.114}$$

(cf. Definition 1.4.7). The fact that O is an orthogonal matrix and the fact that $A=O\Lambda O^*$ therefore demonstrate that for all $v\in\mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}\langle v,Av\rangle&=\langle v,O\Lambda O^*v\rangle=\langle O^*v,\Lambda O^*v\rangle\\&\leq\max(\mathcal{S})\|O^*v\|_2^2=\max(\mathcal{S})\|v\|_2^2.\end{aligned} \tag{6.115}$$

Moreover, observe that (6.107) implies that for all $v=(v_1,v_2,\ldots,v_{\mathfrak{d}})\in\mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}\langle v,\Lambda v\rangle&=\sum_{i=1}^{\mathfrak{d}}\Lambda_{i,i}|v_i|^2\geq\sum_{i=1}^{\mathfrak{d}}\min\{\Lambda_{1,1},\ldots,\Lambda_{\mathfrak{d},\mathfrak{d}}\}|v_i|^2\\&=\min\{\Lambda_{1,1},\ldots,\Lambda_{\mathfrak{d},\mathfrak{d}}\}\|v\|_2^2=\min(\mathcal{S})\|v\|_2^2.\end{aligned} \tag{6.116}$$

The fact that O is an orthogonal matrix and the fact that $A=O\Lambda O^*$ hence demonstrate that for all $v\in\mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}\langle v,Av\rangle&=\langle v,O\Lambda O^*v\rangle=\langle O^*v,\Lambda O^*v\rangle\\&\geq\min(\mathcal{S})\|O^*v\|_2^2=\min(\mathcal{S})\|v\|_2^2.\end{aligned} \tag{6.117}$$

Combining this with (6.115) establishes item (iii). The proof of Lemma 6.1.16 is thus complete. □

We now present the promised Proposition 6.1.17 which discloses suitable conditions (cf. (6.118) and (6.119) below) on the Hessians of the objective function of the considered optimization problem which are sufficient to ensure that (6.46) is satisfied so that we are in the position to apply the error analysis in Sections 6.1.4.1, 6.1.4.2, 6.1.4.3, and 6.1.4.4 above.

Proposition 6.1.17 (Conditions on the spectrum of the Hessian of the objective function at a local minimum point)**.** *Let* $\mathfrak{d} \in \mathbb{N}$*, let* $|||\cdot|||\colon \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}} \to [0,\infty)$ *satisfy for all* $A \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}}$ *that* $|||A||| = \sup_{v\in\mathbb{R}^{\mathfrak{d}}\setminus\{0\}} \frac{\|Av\|_2}{\|v\|_2}$*, and let* $\lambda, \alpha \in (0,\infty)$*,* $\beta \in [\alpha,\infty)$*,* $\vartheta \in \mathbb{R}^{\mathfrak{d}}$*,* $\mathscr{L} \in C^2(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$ *satisfy for all* $v, w \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$(\nabla\mathscr{L})(\vartheta) = 0, \qquad |||(\operatorname{Hess}\mathscr{L})(v) - (\operatorname{Hess}\mathscr{L})(w)||| \leq \lambda\|v-w\|_2, \tag{6.118}$$

$$\text{and} \qquad \{\mu\in\mathbb{R}\colon (\exists\, u\in\mathbb{R}^{\mathfrak{d}}\setminus\{0\}\colon [(\operatorname{Hess}\mathscr{L})(\vartheta)]u = \mu u)\} \subseteq [\alpha,\beta] \tag{6.119}$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \{w\in\mathbb{R}^{\mathfrak{d}}\colon \|w-\vartheta\|_2 \leq \frac{\alpha}{\lambda}\}$ *that*

$$\langle \theta-\vartheta, (\nabla\mathscr{L})(\theta)\rangle \geq \tfrac{\alpha}{2}\|\theta-\vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla\mathscr{L})(\theta)\|_2 \leq \tfrac{3\beta}{2}\|\theta-\vartheta\|_2 \tag{6.120}$$

(cf. Definition 1.4.7).

Proof of Proposition 6.1.17. Throughout this proof, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ be the set given by

$$\mathbb{B} = \{w\in\mathbb{R}^{\mathfrak{d}}\colon \|w-\vartheta\|_2 \leq \tfrac{\alpha}{\lambda}\} \tag{6.121}$$

and let $\mathcal{S} \subseteq \mathbb{C}$ be the set given by

$$\mathcal{S} = \{\mu\in\mathbb{C}\colon (\exists\, u\in\mathbb{C}^{\mathfrak{d}}\setminus\{0\}\colon [(\operatorname{Hess}\mathscr{L})(\vartheta)]u = \mu u)\}. \tag{6.122}$$

Note that the fact that $(\operatorname{Hess}\mathscr{L})(\vartheta) \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}}$ is a symmetric matrix, item (i) in Lemma 6.1.16, and (6.119) imply that

$$\mathcal{S} = \{\mu\in\mathbb{R}\colon (\exists\, u\in\mathbb{R}^{\mathfrak{d}}\setminus\{0\}\colon [(\operatorname{Hess}\mathscr{L})(\vartheta)]u = \mu u)\} \subseteq [\alpha,\beta]. \tag{6.123}$$

Next observe that the assumption that $(\nabla\mathscr{L})(\vartheta) = 0$ and the fundamental theorem of

calculus ensure that for all $\theta, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}
&\langle w, (\nabla \mathscr{L})(\theta)\rangle = \langle w, (\nabla \mathscr{L})(\theta) - (\nabla \mathscr{L})(\vartheta)\rangle \\
&= \Big\langle w, [(\nabla \mathscr{L})(\vartheta + t(\theta - \vartheta))]_{t=0}^{t=1} \Big\rangle \\
&= \left\langle w, \int_0^1 [(\operatorname{Hess} \mathscr{L})(\vartheta + t(\theta - \vartheta))](\theta - \vartheta)\, \mathrm{d}t \right\rangle \\
&= \int_0^1 \big\langle w, [(\operatorname{Hess} \mathscr{L})(\vartheta + t(\theta - \vartheta))](\theta - \vartheta)\big\rangle \, \mathrm{d}t \\
&= \big\langle w, [(\operatorname{Hess} \mathscr{L})(\vartheta)](\theta - \vartheta)\big\rangle \\
&\quad + \int_0^1 \big\langle w, \big[(\operatorname{Hess} \mathscr{L})(\vartheta + t(\theta - \vartheta)) - (\operatorname{Hess} \mathscr{L})(\vartheta)\big](\theta - \vartheta)\big\rangle \, \mathrm{d}t
\end{aligned} \tag{6.124}$$

(cf. Definition 1.4.7). The fact that $(\operatorname{Hess} \mathscr{L})(\vartheta) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ is a symmetric matrix, item (iii) in Lemma 6.1.16, and the Cauchy-Schwarz inequality therefore imply that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned}
&\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta)\rangle \\
&\geq \big\langle \theta - \vartheta, [(\operatorname{Hess} \mathscr{L})(\vartheta)](\theta - \vartheta)\big\rangle \\
&\quad - \left| \int_0^1 \big\langle \theta - \vartheta, \big[(\operatorname{Hess} \mathscr{L})(\vartheta + t(\theta - \vartheta)) - (\operatorname{Hess} \mathscr{L})(\vartheta)\big](\theta - \vartheta)\big\rangle \, \mathrm{d}t \right| \\
&\geq \min(\mathcal{S}) \|\theta - \vartheta\|_2^2 \\
&\quad - \int_0^1 \|\theta - \vartheta\|_2 \big\| \big[(\operatorname{Hess} \mathscr{L})(\vartheta + t(\theta - \vartheta)) - (\operatorname{Hess} \mathscr{L})(\vartheta)\big](\theta - \vartheta)\big\|_2 \, \mathrm{d}t.
\end{aligned} \tag{6.125}$$

Combining this with (6.123) and (6.118) shows that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned}
&\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta)\rangle \\
&\geq \alpha \|\theta - \vartheta\|_2^2 \\
&\quad - \int_0^1 \|\theta - \vartheta\|_2 \|(\operatorname{Hess} \mathscr{L})(\vartheta + t(\theta - \vartheta)) - (\operatorname{Hess} \mathscr{L})(\vartheta)\| \|\theta - \vartheta\|_2 \, \mathrm{d}t \\
&\geq \alpha \|\theta - \vartheta\|_2^2 - \left[\int_0^1 \lambda \|\vartheta + t(\theta - \vartheta) - \vartheta\|_2 \, \mathrm{d}t\right] \|\theta - \vartheta\|_2^2 \\
&= \left(\alpha - \left[\int_0^1 t \, \mathrm{d}t\right] \lambda \|\theta - \vartheta\|_2\right) \|\theta - \vartheta\|_2^2 = \big(\alpha - \tfrac{\lambda}{2} \|\theta - \vartheta\|_2\big) \|\theta - \vartheta\|_2^2 \\
&\geq \big(\alpha - \tfrac{\lambda \alpha}{2\lambda}\big) \|\theta - \vartheta\|_2^2 = \tfrac{\alpha}{2} \|\theta - \vartheta\|_2^2.
\end{aligned} \tag{6.126}$$

Moreover, observe that (6.118), (6.123), (6.124), the fact that $(\operatorname{Hess} \mathscr{L})(\vartheta) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ is a symmetric matrix, item (ii) in Lemma 6.1.16, the Cauchy-Schwarz inequality, and the

assumption that $\alpha \leq \beta$ ensure that for all $\theta \in \mathbb{B}$, $w \in \mathbb{R}^{\mathfrak{d}}$ with $\|w\|_2 = 1$ it holds that

$$\begin{aligned}
&\langle w, (\nabla \mathscr{L})(\theta)\rangle \\
&\leq \big|\big\langle w, [(\operatorname{Hess}\mathscr{L})(\vartheta)](\theta-\vartheta)\big\rangle\big| \\
&\quad + \left|\int_0^1 \big\langle w, \big[(\operatorname{Hess}\mathscr{L})(\vartheta + t(\theta-\vartheta)) - (\operatorname{Hess}\mathscr{L})(\vartheta)\big](\theta-\vartheta)\big\rangle \,\mathrm{d}t\right| \\
&\leq \|w\|_2 \|[(\operatorname{Hess}\mathscr{L})(\vartheta)](\theta-\vartheta)\|_2 \\
&\quad + \int_0^1 \|w\|_2 \|[(\operatorname{Hess}\mathscr{L})(\vartheta + t(\theta-\vartheta)) - (\operatorname{Hess}\mathscr{L})(\vartheta)](\theta-\vartheta)\|_2 \,\mathrm{d}t \\
&\leq \left[\sup_{v\in\mathbb{R}^{\mathfrak{d}}\backslash\{0\}} \frac{\|[(\operatorname{Hess}\mathscr{L})(\vartheta)]v\|_2}{\|v\|_2}\right]\|\theta-\vartheta\|_2 \\
&\quad + \int_0^1 |||(\operatorname{Hess}\mathscr{L})(\vartheta + t(\theta-\vartheta)) - (\operatorname{Hess}\mathscr{L})(\vartheta)|||\,\|\theta-\vartheta\|_2 \,\mathrm{d}t \\
&\leq \max\big(\mathcal{S}\big)\|\theta-\vartheta\|_2 + \left[\int_0^1 \lambda\|\vartheta + t(\theta-\vartheta) - \vartheta\|_2 \,\mathrm{d}t\right]\|\theta-\vartheta\|_2 \\
&\leq \left(\beta + \lambda\left[\int_0^1 t\,\mathrm{d}t\right]\|\theta-\vartheta\|_2\right)\|\theta-\vartheta\|_2 = \big(\beta + \tfrac{\lambda}{2}\|\theta-\vartheta\|_2\big)\|\theta-\vartheta\|_2 \\
&\leq \big(\beta + \tfrac{\lambda\alpha}{2\lambda}\big)\|\theta-\vartheta\|_2 = \big[\tfrac{2\beta+\alpha}{2}\big]\|\theta-\vartheta\|_2 \leq \tfrac{3\beta}{2}\|\theta-\vartheta\|_2.
\end{aligned} \tag{6.127}$$

Therefore, we obtain for all $\theta \in \mathbb{B}$ that

$$\|(\nabla\mathscr{L})(\theta)\|_2 = \sup_{w\in\mathbb{R}^{\mathfrak{d}},\,\|w\|_2=1} [\langle w, (\nabla\mathscr{L})(\theta)\rangle] \leq \tfrac{3\beta}{2}\|\theta-\vartheta\|_2. \tag{6.128}$$

Combining this and (6.126) establishes (6.120). The proof of Proposition 6.1.17 is thus complete. □

The next result, Corollary 6.1.18 below, combines Proposition 6.1.17 with Proposition 6.1.10 to obtain an error analysis which assumes the conditions in (6.118) and (6.119) in Proposition 6.1.17 above. A result similar to Corollary 6.1.18 can, for example, be found in Nesterov [316, Theorem 1.2.4].

Corollary 6.1.18 (Error analysis for the GD optimization method under conditions on the Hessian of the objective function)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $|||\cdot|||\colon \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}} \to \mathbb{R}$ satisfy for all $A \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}}$ that $|||A||| = \sup_{v\in\mathbb{R}^{\mathfrak{d}}\backslash\{0\}} \frac{\|Av\|_2}{\|v\|_2}$, and let $\lambda, \alpha \in (0,\infty)$, $\beta \in [\alpha,\infty)$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0, \frac{4\alpha}{9\beta^2}]$, $\vartheta, \xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$(\nabla\mathscr{L})(\vartheta) = 0, \qquad |||(\operatorname{Hess}\mathscr{L})(v) - (\operatorname{Hess}\mathscr{L})(w)||| \leq \lambda\|v-w\|_2, \tag{6.129}$$

$$\{\mu \in \mathbb{R}\colon (\exists\, u \in \mathbb{R}^{\mathfrak{d}}\backslash\{0\}\colon [(\operatorname{Hess}\mathscr{L})(\vartheta)]u = \mu u)\} \subseteq [\alpha,\beta], \tag{6.130}$$

and $\|\xi - \vartheta\|_2 \leq \frac{\alpha}{\lambda}$, *and let* $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.131}$$

(cf. Definition 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{B} \colon \mathscr{L}(\theta) = \inf_{w \in \mathbb{B}} \mathscr{L}(w)\} = \{\vartheta\}$,

(ii) it holds for all $k \in \mathbb{N}$ *that* $0 \leq 1 - \alpha\gamma_k + \frac{9\beta^2(\gamma_k)^2}{4} \leq 1$,

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

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^{n} \left[1 - \alpha\gamma_k + \tfrac{9\beta^2(\gamma_k)^2}{4}\right]^{1/2}\right] \|\xi - \vartheta\|_2, \tag{6.132}$$

and

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

$$0 \leq \mathscr{L}(\Theta_n) - \mathscr{L}(\vartheta) \leq \tfrac{3\beta}{4} \left[\prod_{k=1}^{n} \left[1 - \alpha\gamma_k + \tfrac{9\beta^2(\gamma_k)^2}{4}\right]\right] \|\xi - \vartheta\|_2^2. \tag{6.133}$$

Proof of Corollary 6.1.18. Note that (6.129), (6.130), and Proposition 6.1.17 prove that for all $\theta \in \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq \frac{\alpha}{\lambda}\}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq \tfrac{\alpha}{2} \|\theta - \vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq \tfrac{3\beta}{2} \|\theta - \vartheta\|_2 \tag{6.134}$$

(cf. Definition 1.4.7). Combining this, the assumption that

$$\|\xi - \vartheta\|_2 \leq \frac{\alpha}{\lambda}, \tag{6.135}$$

(6.131), and items (iv) and (v) in Proposition 6.1.10 (applied with $c \curvearrowleft \frac{\alpha}{2}$, $L \curvearrowleft \frac{3\beta}{2}$, $r \curvearrowleft \frac{\alpha}{\lambda}$ in the notation of Proposition 6.1.10) establishes items (i), (ii), (iii), and (iv). The proof of Corollary 6.1.18 is thus complete. □

Remark 6.1.19. In Corollary 6.1.18 we establish convergence of the considered GD process under, amongst other things, the assumption that all eigenvalues of the Hessian of $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ at the local minimum point ϑ are strictly positive (see (6.130)). In the situation where $\mathcal{L}$ is the cost function (integrated loss function) associated to a supervised learning problem in the training of ANNs, this assumption is basically not satisfied. Nonetheless, the convergence analysis in Corollary 6.1.18 can, roughly speaking, also be performed under the essentially (up to the smoothness conditions) more general assumption that there exists $k \in \mathbb{N}_0$ such that the set of local minimum points is locally a smooth k-dimensional submanifold of

$\mathbb{R}^{\mathfrak{d}}$ and that the rank of the Hessian of $\mathscr{L}$ is on this set of local minimum points locally (at least) $\mathfrak{d} - k$ (cf. Fehrman et al. [137] for details). In certain situations this essentially generalized assumption has also been shown to be satisfied in the training of ANNs in suitable supervised learning problems (see Jentzen & Riekert [232]).

6.1.4.6 Equivalent conditions on the objective function

Lemma 6.1.20. *Let $\mathfrak{d} \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle \colon \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a scalar product, let $|||\cdot||| \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that $|||v||| = \sqrt{\langle\langle v, v \rangle\rangle}$, let $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon |||w - \vartheta||| \leq r\}$, and let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that*

$$|||\theta - \gamma \mathscr{G}(\theta) - \vartheta||| \leq \varepsilon |||\theta - \vartheta|||. \tag{6.136}$$

Then it holds for all $\theta \in \mathbb{B}$ that

$$\begin{aligned} \langle\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\rangle &\geq \max\left\{ \left[\tfrac{1-\varepsilon^2}{2\gamma} \right] |||\theta - \vartheta|||^2, \tfrac{\gamma}{2} |||\mathscr{G}(\theta)|||^2 \right\} \\ &\geq \min\left\{ \tfrac{1-\varepsilon^2}{2\gamma}, \tfrac{\gamma}{2} \right\} \max\left\{ |||\theta - \vartheta|||^2, |||\mathscr{G}(\theta)|||^2 \right\}. \end{aligned} \tag{6.137}$$

Proof of Lemma 6.1.20. First, note that (6.136) ensures that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \varepsilon^2 |||\theta - \vartheta|||^2 &\geq |||\theta - \gamma \mathscr{G}(\theta) - \vartheta|||^2 = |||(\theta - \vartheta) - \gamma \mathscr{G}(\theta)|||^2 \\ &= |||\theta - \vartheta|||^2 - 2\gamma \langle\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\rangle + \gamma^2 |||\mathscr{G}(\theta)|||^2. \end{aligned} \tag{6.138}$$

Hence, we obtain for all $\theta \in \mathbb{B}$ that

$$\begin{aligned} 2\gamma \langle\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\rangle &\geq (1 - \varepsilon^2) |||\theta - \vartheta|||^2 + \gamma^2 |||\mathscr{G}(\theta)|||^2 \\ &\geq \max\left\{ (1 - \varepsilon^2) |||\theta - \vartheta|||^2, \gamma^2 |||\mathscr{G}(\theta)|||^2 \right\} \geq 0. \end{aligned} \tag{6.139}$$

This demonstrates that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \langle\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\rangle &\geq \tfrac{1}{2\gamma} \max\left\{ (1 - \varepsilon^2) |||\theta - \vartheta|||^2, \gamma^2 |||\mathscr{G}(\theta)|||^2 \right\} \\ &= \max\left\{ \left[\tfrac{1-\varepsilon^2}{2\gamma} \right] |||\theta - \vartheta|||^2, \tfrac{\gamma}{2} |||\mathscr{G}(\theta)|||^2 \right\} \\ &\geq \min\left\{ \tfrac{1-\varepsilon^2}{2\gamma}, \tfrac{\gamma}{2} \right\} \max\left\{ |||\theta - \vartheta|||^2, |||\mathscr{G}(\theta)|||^2 \right\}. \end{aligned} \tag{6.140}$$

The proof of Lemma 6.1.20 is thus complete. □

Lemma 6.1.21. *Let $\mathfrak{d} \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle \colon \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be a scalar product, let $|||\cdot||| \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that $|||v||| = \sqrt{\langle\langle v, v \rangle\rangle}$, let $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$,*

$\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|\!|w - \vartheta|\!\| \leq r\}$, *and let* $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $\theta \in \mathbb{B}$ *that*

$$\langle\!\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\!\rangle \geq c \max\{\|\!|\theta - \vartheta|\!\|^2, \|\!|\mathscr{G}(\theta)|\!\|^2\}. \tag{6.141}$$

Then it holds for all $\theta \in \mathbb{B}$ *that*

$$\langle\!\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\!\rangle \geq c\|\!|\theta - \vartheta|\!\|^2 \qquad \text{and} \qquad \|\!|\mathscr{G}(\theta)|\!\| \leq \tfrac{1}{c}\|\!|\theta - \vartheta|\!\|. \tag{6.142}$$

Proof of Lemma 6.1.21. Observe that (6.141) and the Cauchy-Schwarz inequality assure that for all $\theta \in \mathbb{B}$ it holds that

$$\|\!|\mathscr{G}(\theta)|\!\|^2 \leq \max\{\|\!|\theta - \vartheta|\!\|^2, \|\!|\mathscr{G}(\theta)|\!\|^2\} \leq \tfrac{1}{c}\langle\!\langle \theta - \vartheta, \mathscr{G}(\theta) \rangle\!\rangle \leq \tfrac{1}{c}\|\!|\theta - \vartheta|\!\| \, \|\!|\mathscr{G}(\theta)|\!\|. \tag{6.143}$$

Therefore, we obtain for all $\theta \in \mathbb{B}$ that

$$\|\!|\mathscr{G}(\theta)|\!\| \leq \tfrac{1}{c}\|\!|\theta - \vartheta|\!\|. \tag{6.144}$$

Combining this with (6.141) completes the proof of Lemma 6.1.21. □

Lemma 6.1.22. *Let* $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *satisfy for all* $\theta \in \mathbb{B}$ *that*

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2. \tag{6.145}$$

Then it holds for all $v \in \mathbb{R}^{\mathfrak{d}}$, $s, t \in [0, 1]$ *with* $\|v\|_2 \leq r$ *and* $s \leq t$ *that*

$$\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta + sv) \geq \tfrac{c}{2}(t^2 - s^2)\|v\|_2^2. \tag{6.146}$$

Proof of Lemma 6.1.22. First of all, observe that (6.145) implies that for all $v \in \mathbb{R}^{\mathfrak{d}}$ with $\|v\|_2 \leq r$ it holds that

$$\langle (\nabla \mathscr{L})(\vartheta + v), v \rangle \geq c\|v\|_2^2. \tag{6.147}$$

The fundamental theorem of calculus hence ensures that for all $v \in \mathbb{R}^{\mathfrak{d}}$, $s, t \in [0, 1]$ with $\|v\|_2 \leq r$ and $s \leq t$ it holds that

$$\begin{aligned} \mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta + sv) &= \big[\mathscr{L}(\vartheta + hv)\big]_{h=s}^{h=t} \\ &= \int_s^t \mathscr{L}'(\vartheta + hv) v \, \mathrm{d}h \\ &= \int_s^t \tfrac{1}{h}\langle (\nabla \mathscr{L})(\vartheta + hv), hv \rangle \, \mathrm{d}h \\ &\geq \int_s^t \tfrac{c}{h}\|hv\|_2^2 \, \mathrm{d}h \\ &= c\left[\int_s^t h \, \mathrm{d}h\right]\|v\|_2^2 = \tfrac{c}{2}(t^2 - s^2)\|v\|_2^2. \end{aligned} \tag{6.148}$$

The proof of Lemma 6.1.22 is thus complete. □

Lemma 6.1.23. *Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0,\infty)$, $r \in (0,\infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$, $s,t \in [0,1]$ with $\|v\|_2 \leq r$ and $s \leq t$ that*

$$\mathscr{L}(\vartheta + tv) - \mathscr{L}(\vartheta + sv) \geq c(t^2 - s^2)\|v\|_2^2 \tag{6.149}$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq 2c\|\theta - \vartheta\|_2^2 \tag{6.150}$$

(cf. Definition 1.4.7).

Proof of Lemma 6.1.23. Observe that (6.149) ensures that for all $s \in (0,r] \cap \mathbb{R}$, $\theta \in \mathbb{R}^{\mathfrak{d}} \backslash \{\vartheta\}$ with $\|\theta - \vartheta\|_2 < s$ it holds that

$$\begin{aligned}
&\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle = \mathscr{L}'(\theta)(\theta - \vartheta) = \limsup_{h \searrow 0} \left(\tfrac{1}{h} \left[\mathscr{L}(\theta + h(\theta - \vartheta)) - \mathscr{L}(\theta) \right] \right) \\
&= \limsup_{h \searrow 0} \left(\frac{1}{h} \left[\mathscr{L}\left(\vartheta + \tfrac{(1+h)\|\theta - \vartheta\|_2}{s} \left(\tfrac{s}{\|\theta - \vartheta\|_2} (\theta - \vartheta) \right) \right) \right. \right. \\
&\quad \left. \left. - \mathscr{L}\left(\vartheta + \tfrac{\|\theta - \vartheta\|_2}{s} \left(\tfrac{s}{\|\theta - \vartheta\|_2} (\theta - \vartheta) \right) \right) \right] \right) \\
&\geq \limsup_{h \searrow 0} \left(\frac{c}{h} \left(\left[\tfrac{(1+h)\|\theta - \vartheta\|_2}{s} \right]^2 - \left[\tfrac{\|\theta - \vartheta\|_2}{s} \right]^2 \right) \left\| \tfrac{s}{\|\theta - \vartheta\|_2} (\theta - \vartheta) \right\|_2^2 \right) \\
&= c \left[\limsup_{h \searrow 0} \left(\tfrac{(1+h)^2 - 1}{h} \right) \right] \left[\tfrac{\|\theta - \vartheta\|_2}{s} \right]^2 \left\| \tfrac{s}{\|\theta - \vartheta\|_2} (\theta - \vartheta) \right\|_2^2 \\
&= c \left[\limsup_{h \searrow 0} \left(\tfrac{2h + h^2}{h} \right) \right] \|\theta - \vartheta\|_2^2 \\
&= c \left[\limsup_{h \searrow 0} (2 + h) \right] \|\theta - \vartheta\|_2^2 = 2c\|\theta - \vartheta\|_2^2
\end{aligned} \tag{6.151}$$

(cf. Definition 1.4.7). Hence, we obtain that for all $\theta \in \mathbb{R}^{\mathfrak{d}} \backslash \{\vartheta\}$ with $\|\theta - \vartheta\|_2 < r$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq 2c\|\theta - \vartheta\|_2^2. \tag{6.152}$$

Combining this with the fact that the function

$$\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathscr{L})(v) \in \mathbb{R}^{\mathfrak{d}} \tag{6.153}$$

is continuous establishes (6.150). The proof of Lemma 6.1.23 is thus complete. □

Lemma 6.1.24. *Let $\mathfrak{d} \in \mathbb{N}$, $L \in (0,\infty)$, $r \in (0,\infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2 \tag{6.154}$$

(cf. Definition 3.3.4). Then it holds for all $v, w \in \mathbb{B}$ that

$$|\mathscr{L}(v) - \mathscr{L}(w)| \leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\}\|v - w\|_2. \tag{6.155}$$

Proof of Lemma 6.1.24. Observe that (6.154), the fundamental theorem of calculus, and the Cauchy-Schwarz inequality assure that for all $v, w \in \mathbb{B}$ it holds that

$$\begin{aligned}
|\mathscr{L}(v) - \mathscr{L}(w)| &= \Big|\big[\mathscr{L}(w + h(v-w))\big]_{h=0}^{h=1}\Big| \\
&= \left|\int_0^1 \mathscr{L}'(w + h(v-w))(v-w)\,\mathrm{d}h\right| \\
&= \left|\int_0^1 \big\langle (\nabla \mathscr{L})\big(w + h(v-w)\big), v - w\big\rangle\,\mathrm{d}h\right| \\
&\leq \int_0^1 \|(\nabla \mathscr{L})\big(hv + (1-h)w\big)\|_2 \|v - w\|_2\,\mathrm{d}h \\
&\leq \int_0^1 L\|hv + (1-h)w - \vartheta\|_2 \|v - w\|_2\,\mathrm{d}h \\
&\leq \int_0^1 L\big(h\|v - \vartheta\|_2 + (1-h)\|w - \vartheta\|_2\big)\|v - w\|_2\,\mathrm{d}h \\
&= L\,\|v - w\|_2 \left[\int_0^1 \big(h\|v - \vartheta\|_2 + h\|w - \vartheta\|_2\big)\,\mathrm{d}h\right] \\
&= L\,\big(\|v - \vartheta\|_2 + \|w - \vartheta\|_2\big)\,\|v - w\|_2 \left[\int_0^1 h\,\mathrm{d}h\right] \\
&\leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\}\|v - w\|_2
\end{aligned} \tag{6.156}$$

(cf. Definition 1.4.7). The proof of Lemma 6.1.24 is thus complete. □

Lemma 6.1.25. *Let $\mathfrak{d} \in \mathbb{N}$, $L \in (0,\infty)$, $r \in (0,\infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v, w \in \mathbb{B}$ that*

$$|\mathscr{L}(v) - \mathscr{L}(w)| \leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\}\|v - w\|_2 \tag{6.157}$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$\|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2. \tag{6.158}$$

Proof of Lemma 6.1.25. Note that (6.157) implies that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\|_2 < r$ it holds that

$$\begin{aligned}
\|(\nabla \mathscr{L})(\theta)\|_2 &= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2 = 1} \Big[\mathscr{L}'(\theta)(w) \Big] \\
&= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2 = 1} \Big[\limsup_{h \searrow 0} \big[\tfrac{1}{h} (\mathscr{L}(\theta + hw) - \mathscr{L}(\theta)) \big] \Big] \\
&\leq \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2 = 1} \left[\limsup_{h \searrow 0} \left[\tfrac{L}{h} \max\big\{ \|\theta + hw - \vartheta\|_2, \|\theta - \vartheta\|_2 \big\} \|\theta + hw - \theta\|_2 \right] \right] \\
&= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2 = 1} \left[\limsup_{h \searrow 0} \left[L \max\big\{ \|\theta + hw - \vartheta\|_2, \|\theta - \vartheta\|_2 \big\} \tfrac{1}{h} \|hw\|_2 \right] \right] \\
&= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2 = 1} \left[\limsup_{h \searrow 0} \left[L \max\big\{ \|\theta + hw - \vartheta\|_2, \|\theta - \vartheta\|_2 \big\} \right] \right] \\
&= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2 = 1} \Big[L\|\theta - \vartheta\|_2 \Big] = L\|\theta - \vartheta\|_2.
\end{aligned} \tag{6.159}$$

The fact that the function $\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathscr{L})(v) \in \mathbb{R}^{\mathfrak{d}}$ is continuous therefore establishes (6.158). The proof of Lemma 6.1.25 is thus complete. □

Corollary 6.1.26. *Let $\mathfrak{d} \in \mathbb{N}$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} \colon \|w - \vartheta\|_2 \leq r\}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ (cf. Definition 3.3.4). Then the following four statements are equivalent:*

(i) There exist $c, L \in (0, \infty)$ such that for all $\theta \in \mathbb{B}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2 \qquad \text{and} \qquad \|(\nabla \mathscr{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2. \tag{6.160}$$

(ii) There exist $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$ such that for all $\theta \in \mathbb{B}$ it holds that

$$\|\theta - \gamma(\nabla \mathscr{L})(\theta) - \vartheta\|_2 \leq \varepsilon\|\theta - \vartheta\|_2. \tag{6.161}$$

(iii) There exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{B}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathscr{L})(\theta) \rangle \geq c \max\big\{ \|\theta - \vartheta\|_2^2, \|(\nabla \mathscr{L})(\theta)\|_2^2 \big\}. \tag{6.162}$$

(iv) There exist $c, L \in (0,\infty)$ such that for all $v, w \in \mathbb{B}$, $s, t \in [0,1]$ with $s \leq t$ it holds that

$$\mathscr{L}\big(\vartheta + t(v-\vartheta)\big) - \mathscr{L}\big(\vartheta + s(v-\vartheta)\big) \geq c(t^2 - s^2)\|v-\vartheta\|_2^2 \tag{6.163}$$

and

$$|\mathscr{L}(v) - \mathscr{L}(w)| \leq L \max\big\{\|v-\vartheta\|_2, \|w-\vartheta\|_2\big\}\|v-w\|_2 \tag{6.164}$$

(cf. Definition 1.4.7).

Proof of Corollary 6.1.26. Note that items (ii) and (iii) in Lemma 6.1.9 prove that ((i) → (ii)). Observe that Lemma 6.1.20 demonstrates that ((ii) → (iii)). Note that Lemma 6.1.21 establishes that ((iii) → (i)). Observe that Lemma 6.1.22 and Lemma 6.1.24 show that ((i) → (iv)). Note that Lemma 6.1.23 and Lemma 6.1.25 establish that ((iv) → (i)). The proof of Corollary 6.1.26 is thus complete. □

6.2 Explicit midpoint GD optimization

As discussed in Section 6.1 above, the GD optimization method can be viewed as an Euler discretization of the associated GF ODE in Theorem 5.9.4 in Chapter 5. In the literature also more sophisticated methods than the Euler method have been employed to approximate the GF ODE. In particular, higher order Runge-Kutta methods have been used to approximate local minimum points of optimization problems (cf., for instance, Zhang et al. [447]). In this section we illustrate this in the case of the explicit midpoint method.

Definition 6.2.1 (Explicit midpoint GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be differentiable, let $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the explicit midpoint GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ if and only if it holds for all $n \in \mathbb{N}$ that*

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla\mathscr{L})\big(\Theta_{n-1} - \tfrac{\gamma_n}{2}(\nabla\mathscr{L})(\Theta_{n-1})\big). \tag{6.165}$$

Algorithm 6.2.2: Explicit midpoint GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the explicit midpoint GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$ and initial value ξ (cf. Definition 6.2.1)

1: **Initialization:** $\Theta \leftarrow \xi$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \Theta \leftarrow \Theta - \gamma_n(\nabla\mathscr{L})\big(\Theta - \tfrac{\gamma_n}{2}(\nabla\mathscr{L})(\Theta)\big)$
4: **return** Θ

6.2.1 Explicit midpoint discretizations for GF ODEs

Lemma 6.2.3 (Local error of the explicit midpoint method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0,\infty)$, $\mathscr{G} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$, $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $x, y, z \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0,\infty)$ that*

$$\Theta_t = \Theta_0 + \int_0^t \mathscr{G}(\Theta_s)\,\mathrm{d}s, \qquad \theta = \Theta_T + \gamma \mathscr{G}\big(\Theta_T + \tfrac{\gamma}{2}\mathscr{G}(\Theta_T)\big), \tag{6.166}$$

$$\|\mathscr{G}(x)\|_2 \leq c, \qquad \|\mathscr{G}'(x)y\|_2 \leq c\|y\|_2, \qquad \text{and} \qquad \|\mathscr{G}''(x)(y,z)\|_2 \leq c\|y\|_2\|z\|_2 \tag{6.167}$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^3\gamma^3. \tag{6.168}$$

Proof of Lemma 6.2.3. Note that the fundamental theorem of calculus, the assumption that $\mathscr{G} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$, and (6.166) show that for all $t \in [0,\infty)$ it holds that $\Theta \in C^1([0,\infty), \mathbb{R}^{\mathfrak{d}})$ and

$$\dot{\Theta}_t = \mathscr{G}(\Theta_t). \tag{6.169}$$

Combining this with the assumption that $\mathscr{G} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$ 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

$$\ddot{\Theta}_t = \mathscr{G}'(\Theta_t)\dot{\Theta}_t = \mathscr{G}'(\Theta_t)\mathscr{G}(\Theta_t). \tag{6.170}$$

Theorem 6.1.4 and (6.169) hence ensure that

$$\begin{aligned} \Theta_{T+\frac{\gamma}{2}} &= \Theta_T + \left[\frac{\gamma}{2}\right]\dot{\Theta}_T + \int_0^1 (1-r)\left[\frac{\gamma}{2}\right]^2 \ddot{\Theta}_{T+r\gamma/2}\,\mathrm{d}r \\ &= \Theta_T + \left[\frac{\gamma}{2}\right]\mathscr{G}(\Theta_T) + \frac{\gamma^2}{4}\int_0^1 (1-r)\mathscr{G}'(\Theta_{T+r\gamma/2})\mathscr{G}(\Theta_{T+r\gamma/2})\,\mathrm{d}r. \end{aligned} \tag{6.171}$$

Therefore, we obtain that

$$\Theta_{T+\frac{\gamma}{2}} - \Theta_T - \left[\frac{\gamma}{2}\right]\mathscr{G}(\Theta_T) = \frac{\gamma^2}{4}\int_0^1 (1-r)\mathscr{G}'(\Theta_{T+r\gamma/2})\mathscr{G}(\Theta_{T+r\gamma/2})\,\mathrm{d}r. \tag{6.172}$$

Combining this, the fact that for all $x, y \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathscr{G}(x) - \mathscr{G}(y)\|_2 \leq c\|x-y\|_2$, and (6.167) ensures that

$$\begin{aligned} \big\|\mathscr{G}(\Theta_{T+\frac{\gamma}{2}}) - \mathscr{G}\big(\Theta_T + \tfrac{\gamma}{2}\mathscr{G}(\Theta_T)\big)\big\|_2 &\leq c\big\|\Theta_{T+\frac{\gamma}{2}} - \Theta_T - \tfrac{\gamma}{2}\mathscr{G}(\Theta_T)\big\|_2 \\ &\leq \frac{c\gamma^2}{4}\int_0^1 (1-r)\big\|\mathscr{G}'(\Theta_{T+r\gamma/2})\mathscr{G}(\Theta_{T+r\gamma/2})\big\|_2\,\mathrm{d}r \\ &\leq \frac{c^3\gamma^2}{4}\int_0^1 r\,\mathrm{d}r = \frac{c^3\gamma^2}{8}. \end{aligned} \tag{6.173}$$

Furthermore, observe that (6.169), (6.170), the hypothesis that $\mathscr{G} \in C^2(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$, the product rule, and the chain rule establish that for all $t \in [0, \infty)$ it holds that $\Theta \in C^3([0,\infty), \mathbb{R}^\mathfrak{d})$ and

$$\begin{aligned}\dddot{\Theta}_t &= \mathscr{G}''(\Theta_t)(\dot{\Theta}_t, \mathscr{G}(\Theta_t)) + \mathscr{G}'(\Theta_t)\mathscr{G}'(\Theta_t)\dot{\Theta}_t \\ &= \mathscr{G}''(\Theta_t)(\mathscr{G}(\Theta_t), \mathscr{G}(\Theta_t)) + \mathscr{G}'(\Theta_t)\mathscr{G}'(\Theta_t)\mathscr{G}(\Theta_t).\end{aligned} \tag{6.174}$$

Theorem 6.1.4, (6.169), and (6.170) hence imply that for all $s, t \in [0, \infty)$ it holds that

$$\begin{aligned}\Theta_s &= \Theta_t + (s-t)\dot{\Theta}_t + \left[\frac{(s-t)^2}{2}\right]\ddot{\Theta}_t + \int_0^1 \left[\frac{(1-r)^2(s-t)^3}{2}\right]\dddot{\Theta}_{t+r(s-t)}\,\mathrm{d}r \\ &= \Theta_t + (s-t)\mathscr{G}(\Theta_t) + \left[\frac{(s-t)^2}{2}\right]\mathscr{G}'(\Theta_t)\mathscr{G}(\Theta_t) \\ &\quad + \frac{(s-t)^3}{2}\int_0^1 (1-r)^2\big(\mathscr{G}''(\Theta_{t+r(s-t)})(\mathscr{G}(\Theta_{t+r(s-t)}), \mathscr{G}(\Theta_{t+r(s-t)})) \\ &\quad + \mathscr{G}'(\Theta_{t+r(s-t)})\mathscr{G}'(\Theta_{t+r(s-t)})\mathscr{G}(\Theta_{t+r(s-t)})\big)\,\mathrm{d}r.\end{aligned} \tag{6.175}$$

This shows that

$$\begin{aligned}&\Theta_{T+\gamma} - \Theta_T \\ &= \Theta_{T+\frac{\gamma}{2}} + \left[\frac{\gamma}{2}\right]\mathscr{G}(\Theta_{T+\frac{\gamma}{2}}) + \left[\frac{\gamma^2}{8}\right]\mathscr{G}'(\Theta_{T+\frac{\gamma}{2}})\mathscr{G}(\Theta_{T+\frac{\gamma}{2}}) \\ &\quad + \frac{\gamma^3}{16}\int_0^1 (1-r)^2\big(\mathscr{G}''(\Theta_{T+(1+r)\gamma/2})(\mathscr{G}(\Theta_{T+(1+r)\gamma/2}), \mathscr{G}(\Theta_{T+(1+r)\gamma/2})) \\ &\quad + \mathscr{G}'(\Theta_{T+(1+r)\gamma/2})\mathscr{G}'(\Theta_{T+(1+r)\gamma/2})\mathscr{G}(\Theta_{T+(1+r)\gamma/2})\big)\,\mathrm{d}r \\ &\quad - \Bigg[\Theta_{T+\frac{\gamma}{2}} - \left[\frac{\gamma}{2}\right]\mathscr{G}(\Theta_{T+\frac{\gamma}{2}}) + \left[\frac{\gamma^2}{8}\right]\mathscr{G}'(\Theta_{T+\frac{\gamma}{2}})\mathscr{G}(\Theta_{T+\frac{\gamma}{2}}) \\ &\quad - \frac{\gamma^3}{16}\int_0^1 (1-r)^2\big(\mathscr{G}''(\Theta_{T+(1-r)\gamma/2})(\mathscr{G}(\Theta_{T+(1-r)\gamma/2}), \mathscr{G}(\Theta_{T+(1-r)\gamma/2})) \\ &\quad + \mathscr{G}'(\Theta_{T+(1-r)\gamma/2})\mathscr{G}'(\Theta_{T+(1-r)\gamma/2})\mathscr{G}(\Theta_{T+(1-r)\gamma/2})\big)\,\mathrm{d}r\Bigg] \\ &= \gamma\mathscr{G}(\Theta_{T+\frac{\gamma}{2}}) + \frac{\gamma^3}{16}\int_0^1 (1-r)^2\Big(\mathscr{G}''(\Theta_{T+(1+r)\gamma/2})(\mathscr{G}(\Theta_{T+(1+r)\gamma/2}), \mathscr{G}(\Theta_{T+(1+r)\gamma/2})) \\ &\quad + \mathscr{G}'(\Theta_{T+(1+r)\gamma/2})\mathscr{G}'(\Theta_{T+(1+r)\gamma/2})\mathscr{G}(\Theta_{T+(1+r)\gamma/2}) \\ &\quad + \mathscr{G}''(\Theta_{T+(1-r)\gamma/2})(\mathscr{G}(\Theta_{T+(1-r)\gamma/2}), \mathscr{G}(\Theta_{T+(1-r)\gamma/2})) \\ &\quad + \mathscr{G}'(\Theta_{T+(1-r)\gamma/2})\mathscr{G}'(\Theta_{T+(1-r)\gamma/2})\mathscr{G}(\Theta_{T+(1-r)\gamma/2})\Big)\,\mathrm{d}r.\end{aligned} \tag{6.176}$$

This, (6.167), and (6.173) establish that

$$\begin{aligned}\|\Theta_{T+\gamma}-\theta\|_2 &= \big\|\Theta_{T+\gamma}-\Theta_T-\gamma\mathscr{G}(\Theta_T+\tfrac{\gamma}{2}\mathscr{G}(\Theta_T))\big\|_2\\ &\le \big\|\Theta_{T+\gamma}-[\Theta_T+\gamma\mathscr{G}(\Theta_{T+\frac{\gamma}{2}})]\big\|_2+\gamma\big\|\gamma\mathscr{G}(\Theta_{T+\frac{\gamma}{2}})-\mathscr{G}(\Theta_T+\tfrac{\gamma}{2}\mathscr{G}(\Theta_T))\big\|_2\\ &\le \gamma\big\|\mathscr{G}(\Theta_{T+\frac{\gamma}{2}})-\mathscr{G}(\Theta_T+\tfrac{\gamma}{2}\mathscr{G}(\Theta_T))\big\|_2\\ &\quad+\frac{\gamma^3}{16}\int_0^1(1-r)^2\Big(\big\|\mathscr{G}''(\Theta_{T+(1+r)\gamma/2})(\mathscr{G}(\Theta_{T+(1+r)\gamma/2}),\mathscr{G}(\Theta_{T+(1+r)\gamma/2}))\big\|_2\\ &\quad+\big\|\mathscr{G}'(\Theta_{T+(1+r)\gamma/2})\mathscr{G}'(\Theta_{T+(1+r)\gamma/2})\mathscr{G}(\Theta_{T+(1+r)\gamma/2})\big\|_2\\ &\quad+\big\|\mathscr{G}''(\Theta_{T+(1-r)\gamma/2})(\mathscr{G}(\Theta_{T+(1-r)\gamma/2}),\mathscr{G}(\Theta_{T+(1-r)\gamma/2}))\big\|_2\\ &\quad+\big\|\mathscr{G}'(\Theta_{T+(1-r)\gamma/2})\mathscr{G}'(\Theta_{T+(1-r)\gamma/2})\mathscr{G}(\Theta_{T+(1-r)\gamma/2})\big\|_2\Big)\,\mathrm{d}r\\ &\le \frac{c^3\gamma^3}{8}+\frac{c^3\gamma^3}{4}\int_0^1 r^2\,\mathrm{d}r=\frac{5c^3\gamma^3}{24}\le c^3\gamma^3.\end{aligned} \tag{6.177}$$

The proof of Lemma 6.2.3 is thus complete. □

Corollary 6.2.4 (Local error of the explicit midpoint method for GF ODEs)**.** *Let* $\mathfrak{d}\in\mathbb{N}$, $T,\gamma,c\in[0,\infty)$, $\mathscr{L}\in C^3(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$, $\Theta\in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$, $\theta\in\mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $x,y,z\in\mathbb{R}^{\mathfrak{d}}$, $t\in[0,\infty)$ *that*

$$\Theta_t=\Theta_0-\int_0^t(\nabla\mathscr{L})(\Theta_s)\,\mathrm{d}s,\qquad \theta=\Theta_T-\gamma(\nabla\mathscr{L})\big(\Theta_T-\tfrac{\gamma}{2}(\nabla\mathscr{L})(\Theta_T)\big), \tag{6.178}$$

$$\|(\nabla\mathscr{L})(x)\|_2\le c,\quad \|(\operatorname{Hess}\mathscr{L})(x)y\|_2\le c\|y\|_2,\quad \textit{and}\quad \|(\nabla\mathscr{L})''(x)(y,z)\|_2\le c\|y\|_2\|z\|_2 \tag{6.179}$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma}-\theta\|_2\le c^3\gamma^3. \tag{6.180}$$

Proof of Corollary 6.2.4. Throughout this proof, let $\mathscr{G}\colon\mathbb{R}^{\mathfrak{d}}\to\mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta\in\mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{G}(\theta)=-(\nabla\mathscr{L})(\theta). \tag{6.181}$$

Note that the fact that for all $t\in[0,\infty)$ it holds that

$$\Theta_t=\Theta_0+\int_0^t\mathscr{G}(\Theta_s)\,\mathrm{d}s, \tag{6.182}$$

the fact that

$$\theta=\Theta_T+\gamma\mathscr{G}\big(\Theta_T+\tfrac{\gamma}{2}\mathscr{G}(\Theta_T)\big), \tag{6.183}$$

the fact that for all $x \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathscr{G}(x)\|_2 \leq c$, the fact that for all $x, y \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\|\mathscr{G}'(x)y\|_2 \leq c\|y\|_2$, the fact that for all $x, y, z \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\|\mathscr{G}''(x)(y,z)\|_2 \leq c\|y\|_2\|z\|_2, \tag{6.184}$$

and Lemma 6.2.3 demonstrate that

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^3\gamma^3. \tag{6.185}$$

The proof of Corollary 6.2.4 is thus complete. □

6.3 GD optimization with classical momentum

In Section 6.1 above we have introduced and analyzed the classical plain-vanilla GD optimization method. In the literature there are a number of somehow more sophisticated GD-type optimization methods which aim to improve the convergence speed of the classical plain-vanilla GD optimization method (see, for example, Ruder [368] and Sections 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, and 6.11 below). In this section we introduce one of such more sophisticated GD-type optimization methods, that is, we introduce the so-called momentum GD optimization method (see Definition 6.3.1 below). The idea to improve GD optimization methods with a momentum term was first introduced in Polyak [351]. To illustrate the advantage of the momentum GD optimization method over the plain-vanilla GD optimization method we now review a result proving that the momentum GD optimization method does indeed outperform the classical plain-vanilla GD optimization method in the case of a simple class of optimization problems (see Section 6.3.5 below).

In the scientific literature there are several very similar, but not exactly equivalent optimization techniques which are referred to as optimization with momentum. Our definition of the momentum GD optimization method in Definition 6.3.1 below is based on [257, 320] and (7) in [116]. We discuss two alternative definitions from the literature in Section 6.3.1 below and present relationships between these definitions in Section 6.3.2 below.

Definition 6.3.1 (Momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be differentiable, let $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (we say that Θ is the momentum GD process (1^{st} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ) if and only if there exists $\mathbf{m}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \qquad \mathbf{m}_0 = 0, \tag{6.186}$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1-\alpha_n)(\nabla \mathscr{L})(\Theta_{n-1}), \tag{6.187}$$

$$\text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \tag{6.188}$$

Algorithm 6.3.2: Momentum GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the momentum GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (cf. Definition 6.3.1)

1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1-\alpha_n)(\nabla \mathscr{L})(\Theta)$
4: $\quad \Theta \leftarrow \Theta - \gamma_n \mathbf{m}$
5: **return** Θ

Exercise 6.3.1. Let $\mathscr{L}\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that $\mathscr{L}(\theta) = 2\theta^2$ and let Θ be the momentum GD process for the objective function $\mathscr{L}$ with with learning rates $\mathbb{N} \ni n \mapsto 1/2^n \in [0,\infty)$, momentum decay factors $\mathbb{N} \ni n \mapsto 1/2 \in [0,1]$, and initial value 1 (cf. Definition 6.3.1). Specify Θ_1, Θ_2, and Θ_3 explicitly and prove that your results are correct!

Exercise 6.3.2. Let $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$ satisfy $(\xi_1, \xi_2) = (2, 3)$, let $\mathscr{L}\colon \mathbb{R}^2 \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ that

$$\mathscr{L}(\theta) = (\theta_1 - 3)^2 + \tfrac{1}{2}(\theta_2 - 2)^2 + \theta_1 + \theta_2,$$

and let Θ be the momentum GD process for the objective function $\mathscr{L}$ with learning rates $\mathbb{N} \ni n \mapsto 2/n \in [0,\infty)$, momentum decay factors $\mathbb{N} \ni n \mapsto 1/2 \in [0,1]$, and initial value ξ (cf. Definition 6.3.1). Specify Θ_1 and Θ_2 explicitly and prove that your results are correct!

6.3.1 Alternative definitions of GD optimization with momentum

In this section we discuss two definitions similar to the momentum GD optimization method in Definition 6.3.1 which are sometimes also referred to as momentum GD optimization methods in the scientific literature. The method in Definition 6.3.3 below can, for example, be found in [117, Algorithm 2]. The method in Definition 6.3.5 below can, for instance, be found in (9) in [351], (2) in [353], and (4) in [368]. Some relationships between these definitions are discussed in Section 6.3.2 below.

Definition 6.3.3 (Momentum GD optimization method (2^{nd} version))**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be differentiable, let $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let*

$\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *be a function. Then we say that* Θ *is the momentum GD process (2nd version) for the objective function* $\mathscr{L}$ *with learning rates* $(\gamma_n)_{n\in\mathbb{N}}$*, momentum decay factors* $(\alpha_n)_{n\in\mathbb{N}}$*, and initial value* ξ *if and only if there exists* $\mathbf{m}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$\Theta_0 = \xi, \qquad \mathbf{m}_0 = 0, \tag{6.189}$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathscr{L})(\Theta_{n-1}), \tag{6.190}$$

$$\text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \tag{6.191}$$

Algorithm 6.3.4: Momentum GD optimization method (2nd version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the momentum GD process (2nd version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (cf. Definition 6.3.3)

1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathscr{L})(\Theta)$
4: $\quad \Theta \leftarrow \Theta - \gamma_n \mathbf{m}$
5: **return** Θ

Definition 6.3.5 (Momentum GD optimization method (3rd version))**.** *Let* $\mathfrak{d} \in \mathbb{N}$*, let* $\mathscr{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *be differentiable, let* $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$*,* $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$*,* $\xi \in \mathbb{R}^{\mathfrak{d}}$*, and let* $\Theta\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *be a function. Then we say that* Θ *is the momentum GD process (3rd version) for the objective function* $\mathscr{L}$ *with learning rates* $(\gamma_n)_{n\in\mathbb{N}}$*, momentum decay factors* $(\alpha_n)_{n\in\mathbb{N}}$*, and initial value* ξ *if and only if there exists* $\mathbf{m}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$\Theta_0 = \xi, \qquad \mathbf{m}_0 = 0, \tag{6.192}$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1-\alpha_n)\gamma_n (\nabla \mathscr{L})(\Theta_{n-1}), \tag{6.193}$$

$$\text{and} \qquad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \tag{6.194}$$

Algorithm 6.3.6: Momentum GD optimization method (3rd version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the momentum GD process (3rd version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (cf.

Definition 6.3.5)

1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)\gamma_n(\nabla \mathscr{L})(\Theta)$
4: $\quad \Theta \leftarrow \Theta - \mathbf{m}$
5: **return** Θ

Definition 6.3.7 (Momentum GD optimization method (4^{th} version))**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum GD process (4^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ if and only if there exists $\mathbf{m} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \qquad \mathbf{m}_0 = 0, \tag{6.195}$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathscr{L})(\Theta_{n-1}), \tag{6.196}$$

$$\text{and} \qquad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \tag{6.197}$$

Algorithm 6.3.8: Momentum GD optimization method (4^{th} version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the momentum GD process (4^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ (cf. Definition 6.3.7)

1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
2: **for** $n = 1, \ldots, N$ **do**
3: $\quad \mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathscr{L})(\Theta)$
4: $\quad \Theta \leftarrow \Theta - \mathbf{m}$
5: **return** Θ

6.3.2 Relationships between versions of GD optimization with momentum

In this section we discuss relationships between the different versions of the momentum GD optimization method introduced in Definitions 6.3.1, 6.3.3, 6.3.5, and 6.3.7 above.

Proposition 6.3.9 (Comparison of general momentum-type GD optimization methods)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $(\mathfrak{a}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\mathfrak{b}_n^{(1)}\mathfrak{c}_n^{(1)} = \mathfrak{b}_n^{(2)}\mathfrak{c}_n^{(2)} \qquad \text{and} \qquad \frac{\mathfrak{a}_{n+1}^{(1)}\mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\mathfrak{a}_{n+1}^{(2)}\mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}}, \tag{6.198}$$

and for every $i \in \{1,2\}$ *let* $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *and* $\mathbf{m}^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.199}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)}(\nabla\mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.200}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathbf{m}_n^{(i)}. \tag{6.201}$$

Then

$$\Theta^{(1)} = \Theta^{(2)}. \tag{6.202}$$

Proof of Proposition 6.3.9. Throughout this proof, let $\mathscr{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathscr{G}(\theta) = (\nabla\mathscr{L})(\theta). \tag{6.203}$$

Observe that the fact that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{c}_{n+1}^{(1)} = \frac{\mathfrak{c}_{n+1}^{(2)}\mathfrak{b}_{n+1}^{(2)}}{\mathfrak{b}_{n+1}^{(1)}}, \qquad \frac{\mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} = \frac{\mathfrak{b}_{n+1}^{(1)}}{\mathfrak{b}_{n+1}^{(2)}}, \qquad \text{and} \qquad \frac{\mathfrak{b}_{n+1}^{(2)}\mathfrak{a}_{n+1}^{(1)}\mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}\mathfrak{b}_n^{(2)}} = \mathfrak{a}_{n+1}^{(2)} \tag{6.204}$$

proves that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{c}_{n+1}^{(1)}\mathfrak{a}_{n+1}^{(1)}\mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} = \frac{\mathfrak{c}_{n+1}^{(2)}\mathfrak{b}_{n+1}^{(2)}\mathfrak{a}_n^{(1)}\mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}\mathfrak{b}_n^{(2)}} = \mathfrak{c}_{n+1}^{(2)}\mathfrak{a}_{n+1}^{(2)}. \tag{6.205}$$

Furthermore, note that (6.199) implies that

$$\mathbf{m}_0^{(1)} = 0 = \mathbf{m}_0^{(2)} \qquad \text{and} \qquad \Theta_0^{(1)} = \xi = \Theta_0^{(2)}. \tag{6.206}$$

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

$$\mathfrak{c}_n^{(1)}\mathbf{m}_n^{(1)} = \mathfrak{c}_n^{(2)}\mathbf{m}_n^{(2)} \qquad \text{and} \qquad \Theta_n^{(1)} = \Theta_n^{(2)}. \tag{6.207}$$

We now prove (6.207) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ observe that (6.198), (6.199), and (6.206) ensure that

$$\begin{aligned} \mathfrak{c}_1^{(1)} \mathbf{m}_1^{(1)} &= \mathfrak{c}_1^{(1)} (\mathfrak{a}_1^{(1)} \mathbf{m}_0^{(1)} + \mathfrak{b}_1^{(1)} \mathscr{G}(\Theta_0^{(1)})) \\ &= \mathfrak{c}_1^{(1)} \mathfrak{b}_1^{(1)} \mathscr{G}(\Theta_0^{(1)}) \\ &= \mathfrak{c}_1^{(2)} \mathfrak{b}_1^{(2)} \mathscr{G}(\Theta_0^{(2)}) \\ &= \mathfrak{c}_1^{(2)} (\mathfrak{a}_1^{(2)} \mathbf{m}_0^{(2)} + \mathfrak{b}_1^{(2)} \mathscr{G}(\Theta_0^{(2)})) \\ &= \mathfrak{c}_1^{(2)} \mathbf{m}_1^{(2)}. \end{aligned} \tag{6.208}$$

This, (6.201), and (6.206) shows

$$\Theta_1^{(1)} = \Theta_0^{(1)} - \mathfrak{c}_1^{(1)} \mathbf{m}_1^{(1)} = \Theta_0^{(2)} - \mathfrak{c}_1^{(2)} \mathbf{m}_1^{(2)} = \Theta_1^{(2)}. \tag{6.209}$$

Combining this and (6.208) establishes (6.207) in the base case $n = 1$. For the induction step $\mathbb{N} \ni n \to n + 1 \in \{2, 3, \ldots\}$ let $n \in \mathbb{N}$ and assume that

$$\mathfrak{c}_n^{(1)} \mathbf{m}_n^{(1)} = \mathfrak{c}_n^{(2)} \mathbf{m}_n^{(2)} \qquad \text{and} \qquad \Theta_n^{(1)} = \Theta_n^{(2)}. \tag{6.210}$$

Note that (6.198), (6.200), (6.205), and (6.210) establish that

$$\begin{aligned} \mathfrak{c}_{n+1}^{(1)} \mathbf{m}_{n+1}^{(1)} &= \mathfrak{c}_{n+1}^{(1)} (\mathfrak{a}_{n+1}^{(1)} \mathbf{m}_n^{(1)} + \mathfrak{b}_{n+1}^{(1)} \mathscr{G}(\Theta_n^{(1)})) \\ &= \frac{\mathfrak{c}_{n+1}^{(1)} \mathfrak{a}_{n+1}^{(1)} \mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} \mathbf{m}_n^{(2)} + \mathfrak{c}_{n+1}^{(1)} \mathfrak{b}_{n+1}^{(1)} \mathscr{G}(\Theta_n^{(2)}) \\ &= \mathfrak{c}_{n+1}^{(2)} \mathfrak{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)} + \mathfrak{c}_{n+1}^{(2)} \mathfrak{b}_{n+1}^{(2)} \mathscr{G}(\Theta_n^{(2)}) \\ &= \mathfrak{c}_{n+1}^{(2)} (\mathfrak{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)} + \mathfrak{b}_{n+1}^{(2)} \mathscr{G}(\Theta_n^{(2)})) \\ &= \mathfrak{c}_{n+1}^{(2)} \mathbf{m}_{n+1}^{(2)}. \end{aligned} \tag{6.211}$$

This, (6.201), and (6.210) demonstrate that

$$\Theta_{n+1}^{(1)} = \Theta_n^{(1)} - \mathfrak{c}_{n+1}^{(1)} \mathbf{m}_{n+1}^{(1)} = \Theta_n^{(2)} - \mathfrak{c}_{n+1}^{(2)} \mathbf{m}_{n+1}^{(2)} = \Theta_{n+1}^{(2)}. \tag{6.212}$$

Induction thus proves (6.207). Combining (6.206) and (6.207) establishes (6.202). The proof of Proposition 6.3.9 is thus complete. □

Corollary 6.3.10 (Comparison of the 1st and 2nd version of the momentum GD optimization method)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\gamma_n^{(1)} (1 - \alpha_n^{(1)}) = \gamma_n^{(2)} \qquad \text{and} \qquad \frac{\alpha_{n+1}^{(1)} (1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \alpha_{n+1}^{(2)}, \tag{6.213}$$

for every $i \in \{1,2\}$ let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n^{(i)})_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n^{(i)})_{n\in\mathbb{N}}$, and initial value ξ (cf. Definitions 6.3.1 and 6.3.3). Then

$$\Theta^{(1)} = \Theta^{(2)}. \tag{6.214}$$

Proof of Corollary 6.3.10. Throughout this proof let $(\mathfrak{a}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \qquad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \qquad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \tag{6.215}$$

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \qquad \mathfrak{b}_n^{(2)} = 1, \qquad \text{and} \qquad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}. \tag{6.216}$$

Observe that (6.186), (6.187), (6.188), (6.189), (6.190), and (6.191) prove that for all $i \in \{1,2\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.217}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.218}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \tag{6.219}$$

Furthermore, note that (6.213), (6.215), and (6.216) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)}) \gamma_n^{(1)} = \gamma_n^{(2)} = \mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)}. \tag{6.220}$$

Moreover, observe that (6.213), (6.215), and (6.216) ensures that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} (1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \alpha_{n+1}^{(2)} = \frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}}. \tag{6.221}$$

Combining this, (6.217), (6.218), (6.219), and (6.220) with Proposition 6.3.9 shows (6.214). The proof of Corollary 6.3.10 is thus complete. □

Lemma 6.3.11 (Comparison of the 1st and 3rd version of the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\gamma_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\alpha_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,1)$, $(\alpha_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \qquad \text{and} \qquad \frac{\gamma_{n+1}^{(1)} \alpha_{n+1}^{(1)}}{\gamma_n^{(1)}} = \alpha_{n+1}^{(3)}, \tag{6.222}$$

for every $i \in \{1,3\}$ let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n^{(i)})_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n^{(i)})_{n\in\mathbb{N}}$,

and initial value ξ (cf. Definitions 6.3.1 and 6.3.5). Then

$$\Theta^{(1)} = \Theta^{(3)}. \tag{6.223}$$

Proof of Lemma 6.3.11. Throughout this proof let $(\mathfrak{a}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \qquad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \qquad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \tag{6.224}$$

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \qquad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \qquad \text{and} \qquad \mathfrak{c}_n^{(3)} = 1. \tag{6.225}$$

Note that (6.186), (6.187), (6.188), (6.192), (6.193), and (6.194) establish that for all $i \in \{1,3\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.226}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)}(\nabla\mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.227}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathbf{m}_n^{(i)}. \tag{6.228}$$

Furthermore, observe that (6.222), (6.224), and (6.225) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)}\mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)})\gamma_n^{(1)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)} = \mathfrak{b}_n^{(3)}\mathfrak{c}_n^{(3)}. \tag{6.229}$$

Moreover, note that (6.222), (6.224), and (6.225) proves that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\frac{\mathfrak{a}_{n+1}^{(1)}\mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} &= \frac{\alpha_{n+1}^{(1)}(1-\alpha_n^{(1)})}{1-\alpha_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)}\gamma_n^{(3)}(1-\alpha_n^{(3)})\gamma_{n+1}^{(1)}}{\gamma_n^{(1)}\gamma_{n+1}^{(3)}(1-\alpha_{n+1}^{(3)})} \\ &= \frac{\alpha_{n+1}^{(3)}\gamma_n^{(3)}(1-\alpha_n^{(3)})}{\gamma_{n+1}^{(3)}(1-\alpha_{n+1}^{(3)})} = \frac{\mathfrak{a}_{n+1}^{(3)}\mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}}.\end{aligned} \tag{6.230}$$

Combining this, (6.226), (6.227), (6.228), and (6.229) with Proposition 6.3.9 implies (6.223). The proof of Lemma 6.3.11 is thus complete. □

Lemma 6.3.12 (Comparison of the 1st and 4th version of the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\gamma_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\alpha_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,1)$, $(\alpha_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(1)}(1-\alpha_n^{(1)}) = \gamma_n^{(4)} \qquad \text{and} \qquad \frac{\gamma_{n+1}^{(1)}\alpha_{n+1}^{(1)}}{\gamma_n^{(1)}} = \alpha_{n+1}^{(4)}, \tag{6.231}$$

for every $i \in \{1,4\}$ let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the

objective function $\mathscr{L}$ with learning rates $(\gamma_n^{(i)})_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n^{(i)})_{n\in\mathbb{N}}$, and initial value ξ (cf. Definitions 6.3.1 and 6.3.5). Then

$$\Theta^{(1)} = \Theta^{(4)}. \tag{6.232}$$

Proof of Lemma 6.3.12. Throughout this proof let $(\mathfrak{a}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(1)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \qquad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \qquad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \tag{6.233}$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \qquad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \qquad \text{and} \qquad \mathfrak{c}_n^{(4)} = 1. \tag{6.234}$$

Observe that (6.186), (6.187), (6.188), (6.195), (6.196), and (6.197) ensure that for all $i \in \{1,4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.235}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.236}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \tag{6.237}$$

Furthermore, note that (6.231), (6.233), and (6.234) shows that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)}) \gamma_n^{(1)} = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)} \mathfrak{c}_n^{(4)}. \tag{6.238}$$

Moreover, observe that (6.231), (6.233), and (6.234) establishes that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)}(1-\alpha_n^{(1)})}{1-\alpha_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} \gamma_n^{(4)} \gamma_{n+1}^{(1)}}{\gamma_n^{(1)} \gamma_{n+1}^{(4)}} = \frac{\alpha_{n+1}^{(4)} \gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)} \mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \tag{6.239}$$

Combining this, (6.235), (6.236), (6.237), and (6.238) with Proposition 6.3.9 demonstrates (6.232). The proof of Lemma 6.3.12 is thus complete. □

Corollary 6.3.13 (Comparison of the 2nd and 3rd version of the momentum SGD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\gamma_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\alpha_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\alpha_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(2)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \qquad \text{and} \qquad \frac{\gamma_{n+1}^{(2)} \alpha_{n+1}^{(2)}}{\gamma_n^{(2)}} = \alpha_{n+1}^{(3)}, \tag{6.240}$$

for every $i \in \{2,3\}$ let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n^{(i)})_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n^{(i)})_{n\in\mathbb{N}}$,

and initial value ξ (cf. Definitions 6.3.3 and 6.3.7). Then

$$\Theta^{(2)} = \Theta^{(3)}. \tag{6.241}$$

Proof of Corollary 6.3.13. Throughout this proof let $(\mathfrak{a}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \qquad \mathfrak{b}_n^{(2)} = 1, \qquad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}, \tag{6.242}$$

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \qquad \mathfrak{b}_n^{(3)} = (1-\alpha_n^{(3)})\gamma_n^{(3)}, \qquad \text{and} \qquad \mathfrak{c}_n^{(3)} = 1. \tag{6.243}$$

Note that (6.189), (6.190), (6.191), (6.192), (6.193), and (6.194) prove that for all $i \in \{2,3\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.244}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)}(\nabla\mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.245}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathbf{m}_n^{(i)}. \tag{6.246}$$

Furthermore, observe that (6.240), (6.242), and (6.243) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(2)}\mathfrak{c}_n^{(2)} = \gamma_n^{(2)} = \gamma_n^{(3)}(1-\alpha_n^{(3)}) = \mathfrak{b}_n^{(3)}\mathfrak{c}_n^{(3)}. \tag{6.247}$$

Moreover, note that (6.240), (6.242), and (6.243) ensures that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(2)}\mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}} = \alpha_{n+1}^{(2)} = \frac{\alpha_{n+1}^{(3)}\gamma_n^{(3)}(1-\alpha_n^{(3)})}{\gamma_{n+1}^{(3)}(1-\alpha_{n+1}^{(3)})} = \frac{\mathfrak{a}_{n+1}^{(3)}\mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}}. \tag{6.248}$$

Combining this, (6.244), (6.245), (6.246), and (6.247) with Proposition 6.3.9 shows (6.241). The proof of Corollary 6.3.13 is thus complete. □

Lemma 6.3.14 (Comparison of the 2nd and 4th version of the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\gamma_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\alpha_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,1)$, $(\alpha_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(2)} = \gamma_n^{(4)} \qquad \textit{and} \qquad \frac{\gamma_{n+1}^{(2)}\alpha_{n+1}^{(2)}}{\gamma_n^{(2)}} = \alpha_{n+1}^{(4)}, \tag{6.249}$$

for every $i \in \{2,4\}$ let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n^{(i)})_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n^{(i)})_{n\in\mathbb{N}}$,

and initial value ξ (cf. Definitions 6.3.3 and 6.3.5). Then

$$\Theta^{(2)} = \Theta^{(4)}. \tag{6.250}$$

Proof of Lemma 6.3.14. Throughout this proof let $(\mathfrak{a}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(2)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \qquad \mathfrak{b}_n^{(2)} = 1, \qquad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}, \tag{6.251}$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \qquad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \qquad \text{and} \qquad \mathfrak{c}_n^{(4)} = 1. \tag{6.252}$$

Observe that (6.189), (6.190), (6.191), (6.195), (6.196), and (6.197) establish that for all $i \in \{2,4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.253}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.254}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \tag{6.255}$$

Furthermore, note that (6.249), (6.251), and (6.252) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)} = \gamma_n^{(2)} = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)} \mathfrak{c}_n^{(4)}. \tag{6.256}$$

Moreover, observe that (6.249), (6.251), and (6.252) proves that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}} = \alpha_{n+1}^{(2)} = \frac{\alpha_{n+1}^{(4)} \gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)} \mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \tag{6.257}$$

Combining this, (6.253), (6.254), (6.255), and (6.256) with Proposition 6.3.9 implies (6.250). The proof of Lemma 6.3.14 is thus complete. □

Corollary 6.3.15 (Comparison of the 3rd and 4th version of the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\gamma_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\alpha_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,1)$, $(\alpha_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \gamma_n^{(4)} \qquad \text{and} \qquad \alpha_{n+1}^{(3)} = \alpha_{n+1}^{(4)}, \tag{6.258}$$

for every $i \in \{3,4\}$ let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function $\mathscr{L}$ with learning rates $(\gamma_n^{(i)})_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n^{(i)})_{n\in\mathbb{N}}$,

and initial value ξ (cf. Definitions 6.3.5 and 6.3.7). Then

$$\Theta^{(3)} = \Theta^{(4)}. \tag{6.259}$$

Proof of Corollary 6.3.15. Throughout this proof let $(\mathfrak{a}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{a}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{b}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(3)})_{n\in\mathbb{N}} \subseteq (0,\infty)$, $(\mathfrak{c}_n^{(4)})_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \qquad \mathfrak{b}_n^{(3)} = (1-\alpha_n^{(3)})\gamma_n^{(3)}, \qquad \mathfrak{c}_n^{(3)} = 1 \tag{6.260}$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \qquad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \qquad \text{and} \qquad \mathfrak{c}_n^{(4)} = 1, \tag{6.261}$$

Note that (6.192), (6.193), (6.194), (6.195), (6.196), and (6.197) ensure that for all $i \in \{3,4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \qquad \mathbf{m}_0^{(i)} = 0, \tag{6.262}$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)}(\nabla\mathscr{L})(\Theta_{n-1}^{(i)}), \tag{6.263}$$

$$\text{and} \qquad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathbf{m}_n^{(i)}. \tag{6.264}$$

Furthermore, observe that (6.258), (6.260), and (6.261) shows that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(3)}\mathfrak{c}_n^{(3)} = \gamma_n^{(3)}(1-\alpha_n^{(3)}) = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)}\mathfrak{c}_n^{(4)}. \tag{6.265}$$

Moreover, note that (6.258), (6.260), and (6.261) establishes that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(3)}\mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}} = \frac{\alpha_{n+1}^{(3)}(1-\alpha_n^{(3)})\gamma_n^{(3)}}{(1-\alpha_{n+1}^{(3)})\gamma_{n+1}^{(3)}} = \frac{\alpha_{n+1}^{(4)}\gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)}\mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \tag{6.266}$$

Combining this, (6.262), (6.263), (6.264), and (6.265) with Proposition 6.3.9 demonstrates (6.259). The proof of Corollary 6.3.15 is thus complete. □

6.3.3 Representations for GD optimization with momentum

In (6.186), (6.187), and (6.188) above the momentum GD optimization method is formulated by means of a one-step recursion. This one-step recursion can efficiently be exploited in an implementation. In Corollary 6.3.18 below we provide a suitable full-history recursive representation for the momentum GD optimization method, which enables us to develop a better intuition for the momentum GD optimization method. Our proof of Corollary 6.3.18 employs the explicit representation of momentum terms in Lemma 6.3.17 below. Our proof of Lemma 6.3.17, in turn, uses an application of the following result.

Lemma 6.3.16. *Let $(\alpha_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ and let $(m_n)_{n\in\mathbb{N}_0} \subseteq \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $m_0 = 0$ and*

$$m_n = \alpha_n m_{n-1} + 1 - \alpha_n. \tag{6.267}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$m_n = 1 - \prod_{k=1}^{n} \alpha_k. \tag{6.268}$$

Proof of Lemma 6.3.16. We prove (6.268) by induction on $n \in \mathbb{N}_0$. For the base case $n = 0$ observe that the assumption that $m_0 = 0$ proves that

$$m_0 = 0 = 1 - \prod_{k=1}^{0} \alpha_k. \tag{6.269}$$

This establishes (6.268) in the base case $n = 0$. For the induction step note that (6.267) shows that for all $n \in \mathbb{N}_0$ with $m_n = 1 - \prod_{k=1}^{n} \alpha_k$ it holds that

$$\begin{aligned} m_{n+1} &= \alpha_{n+1} m_n + 1 - \alpha_{n+1} = \alpha_{n+1}\left[1 - \prod_{k=1}^{n} \alpha_k\right] + 1 - \alpha_{n+1} \\ &= \alpha_{n+1} - \prod_{k=1}^{n+1} \alpha_k + 1 - \alpha_{n+1} = 1 - \prod_{k=1}^{n+1} \alpha_k. \end{aligned} \tag{6.270}$$

Induction hence establishes (6.268). The proof of Lemma 6.3.16 is thus complete. □

Lemma 6.3.17 (An explicit representation of momentum terms)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$, $(a_{n,k})_{(n,k)\in(\mathbb{N}_0)^2} \subseteq \mathbb{R}$, $(\mathscr{G}_n)_{n\in\mathbb{N}_0} \subseteq \mathbb{R}^{\mathfrak{d}}$, $(\mathbf{m}_n)_{n\in\mathbb{N}_0} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \ldots, n-1\}$ that*

$$\mathbf{m}_0 = 0, \quad \mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)\mathscr{G}_{n-1}, \quad \text{and} \quad a_{n,k} = (1 - \alpha_{k+1})\left[\prod_{l=k+2}^{n} \alpha_l\right] \tag{6.271}$$

Then

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

$$\mathbf{m}_n = \sum_{k=0}^{n-1} a_{n,k} \mathscr{G}_k \tag{6.272}$$

and

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

$$\sum_{k=0}^{n-1} a_{n,k} = 1 - \prod_{k=1}^{n} \alpha_k. \tag{6.273}$$

Proof of Lemma 6.3.17. Throughout this proof, let $(m_n)_{n\in\mathbb{N}_0} \subseteq \mathbb{R}$ satisfy for all $n \in \mathbb{N}_0$ that

$$m_n = \sum_{k=0}^{n-1} a_{n,k}. \tag{6.274}$$

We now prove item (i) by induction on $n \in \mathbb{N}_0$. For the base case $n = 0$ note that (6.271) ensures that

$$\mathbf{m}_0 = 0 = \sum_{k=0}^{-1} a_{0,k}\mathscr{G}_k. \tag{6.275}$$

This establishes item (i) in the base case $n = 0$. For the induction step note that (6.271) establishes that for all $n \in \mathbb{N}_0$ with $\mathbf{m}_n = \sum_{k=0}^{n-1} a_{n,k}\mathscr{G}_k$ it holds that

$$\begin{aligned}
\mathbf{m}_{n+1} &= \alpha_{n+1}\mathbf{m}_n + (1-\alpha_{n+1})\mathscr{G}_n \\
&= \left[\sum_{k=0}^{n-1} \alpha_{n+1} a_{n,k}\mathscr{G}_k\right] + (1-\alpha_{n+1})\mathscr{G}_n \\
&= \left[\sum_{k=0}^{n-1} \alpha_{n+1}(1-\alpha_{k+1})\left[\prod_{l=k+2}^{n} \alpha_l\right]\mathscr{G}_k\right] + (1-\alpha_{n+1})\mathscr{G}_n \\
&= \left[\sum_{k=0}^{n-1} (1-\alpha_{k+1})\left[\prod_{l=k+2}^{n+1} \alpha_l\right]\mathscr{G}_k\right] + (1-\alpha_{n+1})\mathscr{G}_n \\
&= \sum_{k=0}^{n} (1-\alpha_{k+1})\left[\prod_{l=k+2}^{n+1} \alpha_l\right]\mathscr{G}_k = \sum_{k=0}^{n} a_{n+1,k}\mathscr{G}_k.
\end{aligned} \tag{6.276}$$

Induction thus proves item (i). Furthermore, observe that (6.271) and (6.274) demonstrate that for all $n \in \mathbb{N}$ it holds that $m_0 = 0$ and

$$\begin{aligned}
m_n &= \sum_{k=0}^{n-1} a_{n,k} = \sum_{k=0}^{n-1} (1-\alpha_{k+1})\left[\prod_{l=k+2}^{n} \alpha_l\right] = 1 - \alpha_n + \sum_{k=0}^{n-2} (1-\alpha_{k+1})\left[\prod_{l=k+2}^{n} \alpha_l\right] \\
&= 1 - \alpha_n + \sum_{k=0}^{n-2} (1-\alpha_{k+1})\alpha_n\left[\prod_{l=k+2}^{n-1} \alpha_l\right] = 1 - \alpha_n + \alpha_n \sum_{k=0}^{n-2} a_{n-1,k} = 1 - \alpha_n + \alpha_n m_{n-1}.
\end{aligned} \tag{6.277}$$

Combining this with Lemma 6.3.16 implies that for all $n \in \mathbb{N}_0$ it holds that

$$m_n = 1 - \prod_{k=1}^{n} \alpha_k. \tag{6.278}$$

This establishes item (ii). The proof of Lemma 6.3.17 is thus complete. □

Corollary 6.3.18 (On a representation of the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1]$, $(a_{n,k})_{(n,k)\in(\mathbb{N}_0)^2} \subseteq \mathbb{R}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0,1,\ldots,n-1\}$ that*

$$a_{n,k} = (1-\alpha_{k+1})\left[\prod_{l=k+2}^{n} \alpha_l\right], \tag{6.279}$$

let $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$, and let Θ be the momentum GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (cf. Definition 6.3.1). Then

(i) it holds for all $n \in \mathbb{N}$, $k \in \{0,1,\ldots,n-1\}$ that $0 \leq a_{n,k} \leq 1$,

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

$$\sum_{k=0}^{n-1} a_{n,k} = 1 - \prod_{k=1}^{n} \alpha_k, \tag{6.280}$$

and

(iii) it holds for all $n \in \mathbb{N}$ that

$$\Theta_n = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathscr{L})(\Theta_k)\right]. \tag{6.281}$$

Proof of Corollary 6.3.18. Throughout this proof, let $\mathbf{m}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\mathbf{m}_0 = 0 \qquad \text{and} \qquad \mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1-\alpha_n)(\nabla \mathscr{L})(\Theta_{n-1}). \tag{6.282}$$

Note that (6.279) establishes item (i). Observe that (6.279), (6.282), and Lemma 6.3.17 show that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbf{m}_n = \sum_{k=0}^{n-1} a_{n,k} (\nabla \mathscr{L})(\Theta_k) \qquad \text{and} \qquad \sum_{k=0}^{n-1} a_{n,k} = 1 - \prod_{k=1}^{n} \alpha_k. \tag{6.283}$$

This proves item (ii). Note that (6.186), (6.187), (6.188), (6.282), and (6.283) ensure that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathscr{L})(\Theta_k) \right]. \tag{6.284}$$

This establishes item (iii). The proof of Corollary 6.3.18 is thus complete. □

6.3.4 Bias-adjusted GD optimization with momentum

Definition 6.3.19 (Bias-adjusted momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be differentiable, let $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the bias-adjusted momentum GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ if and only if there exists $\mathbf{m} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \qquad \mathbf{m}_0 = 0, \tag{6.285}$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1-\alpha_n)(\nabla \mathscr{L})(\Theta_{n-1}), \tag{6.286}$$

$$\text{and} \qquad \Theta_n = \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^{n} \alpha_l}. \tag{6.287}$$

Algorithm 6.3.20: Bias-adjusted momentum GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N-th step of the bias-adjusted momentum GD process for the objective function $\mathscr{L}$ with learning rates $(\gamma_n)_{n\in\mathbb{N}}$, momentum decay factors $(\alpha_n)_{n\in\mathbb{N}}$, and initial value ξ (cf. Definition 6.3.19)

Initialization: $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
for $n = 1, \ldots, N$ **do**
  $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1-\alpha_n)(\nabla \mathscr{L})(\Theta)$
  $\Theta \leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^{n} \alpha_l}$
return Θ

Corollary 6.3.21 (On a representation of the bias-adjusted momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq [0,\infty)$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $(a_{n,k})_{(n,k)\in(\mathbb{N}_0)^2} \subseteq \mathbb{R}$*

satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \ldots, n-1\}$ *that*

$$a_{n,k} = \frac{(1-\alpha_{k+1})\left[\prod_{l=k+2}^{n} \alpha_l\right]}{1 - \prod_{l=1}^{n} \alpha_l}, \tag{6.288}$$

let $\mathscr{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, *and let* Θ *be the bias-adjusted momentum GD process for the objective function* $\mathscr{L}$ *with learning rates* $(\gamma_n)_{n\in\mathbb{N}}$, *momentum decay factors* $(\alpha_n)_{n\in\mathbb{N}}$, *and initial value* ξ *(cf. Definition 6.3.19). Then*

(i) it holds for all $n \in \mathbb{N}$, $k \in \{0, 1, \ldots, n-1\}$ *that* $0 \leq a_{n,k} \leq 1$,

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

$$\sum_{k=0}^{n-1} a_{n,k} = 1, \tag{6.289}$$

and

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

$$\Theta_n = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathscr{L})(\Theta_k)\right]. \tag{6.290}$$

Proof of Corollary 6.3.21. Throughout this proof, let $\mathbf{m} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\mathbf{m}_0 = 0 \qquad \text{and} \qquad \mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1-\alpha_n)(\nabla \mathscr{L})(\Theta_{n-1}) \tag{6.291}$$

and let $(b_{n,k})_{(n,k)\in(\mathbb{N}_0)^2} \subseteq \mathbb{R}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \ldots, n-1\}$ that

$$b_{n,k} = (1-\alpha_{k+1}) \left[\prod_{l=k+2}^{n} \alpha_l\right]. \tag{6.292}$$

Observe that (6.288) implies item (i). Note that (6.288), (6.291), (6.292), and Lemma 6.3.17 establish that for all $n \in \mathbb{N}$ it holds that

$$\mathbf{m}_n = \sum_{k=0}^{n-1} b_{n,k} (\nabla \mathscr{L})(\Theta_k) \qquad \text{and} \qquad \sum_{k=0}^{n-1} a_{n,k} = \frac{\sum_{k=0}^{n-1} b_{n,k}}{1 - \prod_{k=1}^{n} \alpha_k} = \frac{1 - \prod_{k=1}^{n} \alpha_k}{1 - \prod_{k=1}^{n} \alpha_k} = 1. \tag{6.293}$$

This proves item (ii). Observe that (6.285), (6.286), (6.287), (6.288), (6.291), (6.292), and

(6.293) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Theta_n &= \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^{n} \alpha_l} = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} \left[\frac{b_{n,k}}{1 - \prod_{l=1}^{n} \alpha_l} \right] (\nabla \mathscr{L})(\Theta_k) \right] \\ &= \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathscr{L})(\Theta_k) \right]. \end{aligned} \tag{6.294}$$

This establishes item (iii). The proof of Corollary 6.3.21 is thus complete. □

6.3.5 Error analysis for GD optimization with momentum

In this subsection we provide in Section 6.3.5.2 below an error analysis for the momentum GD optimization method in the case of a class of quadratic objective functions (cf. Proposition 6.3.26 in Section 6.3.5.2 for the precise statement). In this specific case we also provide in Section 6.3.5.3 below a comparison of the convergence speeds of the plain-vanilla GD optimization method and the momentum GD optimization method. In particular, we prove, roughly speeking, that the momentum GD optimization method outperfoms the plain-vanilla GD optimization method in the case of the considered class of quadratic objective functions; see Corollary 6.3.28 in Section 6.3.5.3 for the precise statement. For this comparison between the plain-vanilla GD optimization method and the momentum GD optimization method we employ a refined error analysis of the plain-vanilla GD optimization method for the considered class of quadratic objective functions. This refined error analysis is the subject of the next section (Section 6.3.5.1 below).

In the literature similar error analyses for the momentum GD optimization method can, for example, be found in [49, Section 7.1] and [351].

6.3.5.1 Error analysis for GD optimization in the case of quadratic objective functions

Lemma 6.3.22 (Error analysis for the GD optimization method in the case of quadratic objective functions)**.** *Let* $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\kappa, \mathcal{K}, \lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}} \in (0, \infty)$ *satisfy* $\kappa = \min\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$ *and* $\mathcal{K} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$, *let* $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$\mathscr{L}(\theta) = \tfrac{1}{2} \left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2 \right], \tag{6.295}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{2}{\mathcal{K}+\kappa} (\nabla \mathscr{L})(\Theta_{n-1}). \tag{6.296}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n \|\xi - \vartheta\|_2 \tag{6.297}$$

(cf. Definition 3.3.4).

Proof of Lemma 6.3.22. Throughout this proof, let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$, satisfy for all $n \in \mathbb{N}_0$ that $\Theta_n = (\Theta_n^{(1)}, \Theta_n^{(2)}, \ldots, \Theta_n^{(\mathfrak{d})})$. Note that (6.295) implies that for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\left(\tfrac{\partial \mathscr{L}}{\partial \theta_i}\right)(\theta) = \lambda_i(\theta_i - \vartheta_i). \tag{6.298}$$

Combining this and (6.296) ensures that for all $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} - \vartheta_i &= \Theta_{n-1}^{(i)} - \tfrac{2}{\mathcal{K}+\kappa}\left(\tfrac{\partial \mathscr{L}}{\partial \theta_i}\right)(\Theta_{n-1}) - \vartheta_i \\ &= \Theta_{n-1}^{(i)} - \vartheta_i - \tfrac{2}{\mathcal{K}+\kappa}\left[\lambda_i(\Theta_{n-1}^{(i)} - \vartheta_i)\right] \\ &= \left(1 - \tfrac{2\lambda_i}{\mathcal{K}+\kappa}\right)(\Theta_{n-1}^{(i)} - \vartheta_i). \end{aligned} \tag{6.299}$$

Hence, we obtain that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2^2 &= \sum_{i=1}^{\mathfrak{d}} |\Theta_n^{(i)} - \vartheta_i|^2 \\ &= \sum_{i=1}^{\mathfrak{d}} \left[\left|1 - \tfrac{2\lambda_i}{\mathcal{K}+\kappa}\right|^2 |\Theta_{n-1}^{(i)} - \vartheta_i|^2\right] \\ &\leq \left[\max\left\{\left|1 - \tfrac{2\lambda_1}{\mathcal{K}+\kappa}\right|^2, \ldots, \left|1 - \tfrac{2\lambda_{\mathfrak{d}}}{\mathcal{K}+\kappa}\right|^2\right\}\right] \left[\sum_{i=1}^{\mathfrak{d}} |\Theta_{n-1}^{(i)} - \vartheta_i|^2\right] \\ &= \left[\max\left\{\left|1 - \tfrac{2\lambda_1}{\mathcal{K}+\kappa}\right|, \ldots, \left|1 - \tfrac{2\lambda_{\mathfrak{d}}}{\mathcal{K}+\kappa}\right|\right\}\right]^2 \|\Theta_{n-1} - \vartheta\|_2^2 \end{aligned} \tag{6.300}$$

(cf. Definition 3.3.4). Moreover, note that the fact that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that $\lambda_i \geq \kappa$ implies that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$1 - \tfrac{2\lambda_i}{\mathcal{K}+\kappa} \leq 1 - \tfrac{2\kappa}{\mathcal{K}+\kappa} = \tfrac{\mathcal{K}+\kappa-2\kappa}{\mathcal{K}+\kappa} = \tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \geq 0. \tag{6.301}$$

In addition, observe that the fact that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that $\lambda_i \leq \mathcal{K}$ implies that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$1 - \tfrac{2\lambda_i}{\mathcal{K}+\kappa} \geq 1 - \tfrac{2\mathcal{K}}{\mathcal{K}+\kappa} = \tfrac{\mathcal{K}+\kappa-2\mathcal{K}}{\mathcal{K}+\kappa} = -\left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right] \leq 0. \tag{6.302}$$

This and (6.301) ensure that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\left|1 - \tfrac{2\lambda_i}{\mathcal{K}+\kappa}\right| \leq \tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}. \tag{6.303}$$

Combining this with (6.300) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\leq \left[\max\left\{\left|1 - \tfrac{2\lambda_1}{\mathcal{K}+\kappa}\right|, \ldots, \left|1 - \tfrac{2\lambda_{\mathfrak{d}}}{\mathcal{K}+\kappa}\right|\right\}\right] \|\Theta_{n-1} - \vartheta\|_2 \\ &\leq \left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right] \|\Theta_{n-1} - \vartheta\|_2. \end{aligned} \tag{6.304}$$

Induction therefore establishes that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n \|\Theta_0 - \vartheta\|_2 = \left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n \|\xi - \vartheta\|_2. \tag{6.305}$$

The proof of Lemma 6.3.22 is thus complete. □

Lemma 6.3.22 above establishes, roughly speaking, the convergence rate $\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}$ (see (6.297) above for the precise statement) for the GD optimization method in the case of the objective function in (6.295). The next result, Lemma 6.3.23 below, essentially proves in the situation of Lemma 6.3.22 that this convergence rate cannot be improved by means of a difference choice of the learning rate.

Lemma 6.3.23 (Lower bound for the convergence rate of GD for quadratic objective functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\xi = (\xi_1, \ldots, \xi_{\mathfrak{d}})$, $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\gamma, \kappa, \mathcal{K}, \lambda_1, \lambda_2 \ldots, \lambda_{\mathfrak{d}} \in (0, \infty)$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$ and $\mathcal{K} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2 \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \tfrac{1}{2}\left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2\right], \tag{6.306}$$

and let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \gamma(\nabla \mathscr{L})(\Theta_{n-1}). \tag{6.307}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\geq \left[\max\{\gamma\mathcal{K} - 1, 1 - \gamma\kappa\}\right]^n \left[\min\left\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\right\}\right] \\ &\geq \left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n \left[\min\left\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\right\}\right] \end{aligned} \tag{6.308}$$

(cf. Definition 3.3.4).

Proof of Lemma 6.3.23. Throughout this proof, let $\Theta^{(i)} \colon \mathbb{N}_0 \to \mathbb{R}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$, satisfy for all $n \in \mathbb{N}_0$ that $\Theta_n = (\Theta_n^{(1)}, \ldots, \Theta_n^{(\mathfrak{d})})$ and let $\iota, \mathcal{I} \in \{1, 2, \ldots, \mathfrak{d}\}$ satisfy $\lambda_\iota = \kappa$ and $\lambda_{\mathcal{I}} = \mathcal{K}$. Observe that (6.306) implies that for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\left(\tfrac{\partial \mathscr{L}}{\partial \theta_i}\right)(\theta) = \lambda_i(\theta_i - \vartheta_i). \tag{6.309}$$

Combining this with (6.307) implies that for all $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} - \vartheta_i &= \Theta_{n-1}^{(i)} - \gamma\big(\tfrac{\partial \mathscr{L}}{\partial \theta_i}\big)(\Theta_{n-1}) - \vartheta_i \\ &= \Theta_{n-1}^{(i)} - \vartheta_i - \gamma\lambda_i(\Theta_{n-1}^{(i)} - \vartheta_i) \\ &= (1 - \gamma\lambda_i)(\Theta_{n-1}^{(i)} - \vartheta_i). \end{aligned} \tag{6.310}$$

Induction and (6.307) hence prove that for all $n \in \mathbb{N}_0$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\Theta_n^{(i)} - \vartheta_i = (1 - \gamma\lambda_i)^n(\Theta_0^{(i)} - \vartheta_i) = (1 - \gamma\lambda_i)^n(\xi_i - \vartheta_i). \tag{6.311}$$

This shows that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} &\|\Theta_n - \vartheta\|_2^2 = \sum_{i=1}^{\mathfrak{d}} |\Theta_n^{(i)} - \vartheta_i|^2 = \sum_{i=1}^{\mathfrak{d}} \Big[|1 - \gamma\lambda_i|^{2n} |\xi_i - \vartheta_i|^2\Big] \\ &\geq \big[\min\big\{|\xi_1 - \vartheta_1|^2, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|^2\big\}\big] \left[\sum_{i=1}^{\mathfrak{d}} |1 - \gamma\lambda_i|^{2n}\right] \\ &\geq \big[\min\big\{|\xi_1 - \vartheta_1|^2, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|^2\big\}\big] \big[\max\{|1 - \gamma\lambda_1|^{2n}, \ldots, |1 - \gamma\lambda_{\mathfrak{d}}|^{2n}\}\big] \\ &= \big[\min\big\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\big\}\big]^2 \big[\max\{|1 - \gamma\lambda_1|, \ldots, |1 - \gamma\lambda_{\mathfrak{d}}|\}\big]^{2n} \end{aligned} \tag{6.312}$$

(cf. Definition 3.3.4). Furthermore, note that

$$\begin{aligned} &\max\{|1 - \gamma\lambda_1|, \ldots, |1 - \gamma\lambda_{\mathfrak{d}}|\} \geq \max\{|1 - \gamma\lambda_{\mathcal{I}}|, |1 - \gamma\lambda_\iota|\} \\ &= \max\{|1 - \gamma\mathcal{K}|, |1 - \gamma\kappa|\} = \max\{1 - \gamma\mathcal{K}, \gamma\mathcal{K} - 1, 1 - \gamma\kappa, \gamma\kappa - 1\} \\ &= \max\{\gamma\mathcal{K} - 1, 1 - \gamma\kappa\}. \end{aligned} \tag{6.313}$$

In addition, observe that for all $\alpha \in (-\infty, \frac{2}{\mathcal{K}+\kappa}]$ it holds that

$$\max\{\alpha\mathcal{K} - 1, 1 - \alpha\kappa\} \geq 1 - \alpha\kappa \geq 1 - \big[\tfrac{2}{\mathcal{K}+\kappa}\big]\kappa = \tfrac{\mathcal{K}+\kappa-2\kappa}{\mathcal{K}+\kappa} = \tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}. \tag{6.314}$$

Moreover, note that for all $\alpha \in [\frac{2}{\mathcal{K}+\kappa}, \infty)$ it holds that

$$\max\{\alpha\mathcal{K} - 1, 1 - \alpha\kappa\} \geq \alpha\mathcal{K} - 1 \geq \big[\tfrac{2}{\mathcal{K}+\kappa}\big]\mathcal{K} - 1 = \tfrac{2\mathcal{K}-(\mathcal{K}+\kappa)}{\mathcal{K}+\kappa} = \tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}. \tag{6.315}$$

Combining this, (6.313), and (6.314) proves that

$$\max\{|1 - \gamma\lambda_1|, \ldots, |1 - \gamma\lambda_{\mathfrak{d}}|\} \geq \max\{\gamma\mathcal{K} - 1, 1 - \gamma\kappa\} \geq \tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \geq 0. \tag{6.316}$$

This and (6.312) demonstrate that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\geq \big[\max\{|1 - \gamma\lambda_1|, \ldots, |1 - \gamma\lambda_{\mathfrak{d}}|\}\big]^n \big[\min\big\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\big\}\big] \\ &\geq \big[\max\{\gamma\mathcal{K} - 1, 1 - \gamma\kappa\}\big]^n \big[\min\big\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\big\}\big] \\ &\geq \big[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\big]^n \big[\min\big\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\big\}\big]. \end{aligned} \tag{6.317}$$

The proof of Lemma 6.3.23 is thus complete. □

6.3.5.2 Error analysis for GD optimization with momentum in the case of quadratic objective functions

In this subsection we provide in Proposition 6.3.26 below an error analysis for the momentum GD optimization method in the case of a class of quadratic objective functions. Our proof of Proposition 6.3.26 employs the two auxiliary results on quadratic matrices in Lemma 6.3.24 and Lemma 6.3.25 below. Lemma 6.3.24 is a special case of the so-called Gelfand spectral radius formula in the literature. Lemma 6.3.25 establishes a formula for the determinants of quadratic block matrices (see (6.319) below for the precise statement). Lemma 6.3.25 and its proof can, for instance, be found in Silvester [391, Theorem 3].

Lemma 6.3.24 (A special case of Gelfand's spectral radius formula for real matrices)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $A \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}}$, $\mathcal{S} = \{\lambda \in \mathbb{C} \colon (\exists\, v \in \mathbb{C}^{\mathfrak{d}}\backslash\{0\} \colon Av = \lambda v)\}$ and let $\|\!|\cdot|\!\| \colon \mathbb{R}^{\mathfrak{d}} \to [0,\infty)$ be a norm. Then*

$$\liminf_{n\to\infty}\left(\left[\sup_{v\in\mathbb{R}^{\mathfrak{d}}\backslash\{0\}}\frac{\|\!|A^n v|\!\|}{\|\!|v|\!\|}\right]^{1/n}\right) = \limsup_{n\to\infty}\left(\left[\sup_{v\in\mathbb{R}^{\mathfrak{d}}\backslash\{0\}}\frac{\|\!|A^n v|\!\|}{\|\!|v|\!\|}\right]^{1/n}\right) = \max_{\lambda\in\mathcal{S}\cup\{0\}}|\lambda|. \tag{6.318}$$

Proof of Lemma 6.3.24. Note that, for example, Einsiedler & Ward [132, Theorem 11.6] establishes (6.318) (cf., for instance, Tropp [409]). The proof of Lemma 6.3.24 is thus complete. □

Lemma 6.3.25 (Determinants for block matrices)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $A, B, C, D \in \mathbb{C}^{\mathfrak{d}\times\mathfrak{d}}$ satisfy $CD = DC$. Then*

$$\det(\underbrace{\begin{pmatrix} A & B \\ C & D \end{pmatrix}}_{\in\mathbb{R}^{(2\mathfrak{d})\times(2\mathfrak{d})}}) = \det(AD - BC) \tag{6.319}$$

Proof of Lemma 6.3.25. Throughout this proof, let $\mathcal{D}_x \in \mathbb{C}^{\mathfrak{d}\times\mathfrak{d}}$, $x \in \mathbb{C}$, satisfy for all $x \in \mathbb{C}$ that

$$\mathcal{D}_x = D - x\,\mathrm{I}_{\mathfrak{d}} \tag{6.320}$$

(cf. Definition 1.5.5). Observe that the fact that for all $x \in \mathbb{C}$ it holds that $C\mathcal{D}_x = \mathcal{D}_x C$ and the fact that for all $X, Y, Z \in \mathbb{C}^{\mathfrak{d}\times\mathfrak{d}}$ it holds that

$$\det\begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} = \det(X)\det(Z) = \det\begin{pmatrix} X & 0 \\ Y & Z \end{pmatrix} \tag{6.321}$$

(cf., for example, Petersen [345, Proposition 5.5.3 and Proposition 5.5.4]) imply that for all

$x \in \mathbb{C}$ it holds that

$$\begin{aligned}\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\begin{pmatrix} \mathcal{D}_x & 0 \\ -C & \mathrm{I}_{\mathfrak{d}} \end{pmatrix}\right) &= \det\begin{pmatrix} (A\mathcal{D}_x - BC) & B \\ (C\mathcal{D}_x - \mathcal{D}_x C) & \mathcal{D}_x \end{pmatrix} \\ &= \det\begin{pmatrix} (A\mathcal{D}_x - BC) & B \\ 0 & \mathcal{D}_x \end{pmatrix} \\ &= \det(A\mathcal{D}_x - BC)\det(\mathcal{D}_x).\end{aligned} \tag{6.322}$$

Moreover, note that (6.321) and the multiplicative property of the determinant (see, for instance, Petersen [345, (1) in Proposition 5.5.2]) imply that for all $x \in \mathbb{C}$ it holds that

$$\begin{aligned}\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\begin{pmatrix} \mathcal{D}_x & 0 \\ -C & \mathrm{I}_{\mathfrak{d}} \end{pmatrix}\right) &= \det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\det\begin{pmatrix} \mathcal{D}_x & 0 \\ -C & \mathrm{I}_{\mathfrak{d}} \end{pmatrix} \\ &= \det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\det(\mathcal{D}_x)\det(\mathrm{I}_{\mathfrak{d}}) \\ &= \det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\det(\mathcal{D}_x).\end{aligned} \tag{6.323}$$

Combining this and (6.322) demonstrates that for all $x \in \mathbb{C}$ it holds that

$$\det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\det(\mathcal{D}_x) = \det(A\mathcal{D}_x - BC)\det(\mathcal{D}_x). \tag{6.324}$$

Hence, we obtain for all $x \in \mathbb{C}$ that

$$\left(\det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} - \det(A\mathcal{D}_x - BC)\right)\det(\mathcal{D}_x) = 0. \tag{6.325}$$

This implies that for all $x \in \mathbb{C}$ with $\det(\mathcal{D}_x) \neq 0$ it holds that

$$\det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} - \det(A\mathcal{D}_x - BC) = 0. \tag{6.326}$$

Moreover, note that the fact that $\mathbb{C} \ni x \mapsto \det(D - x\,\mathrm{I}_{\mathfrak{d}}) \in \mathbb{C}$ is a polynomial function of degree $\mathfrak{d}$ ensures that $\{x \in \mathbb{C} \colon \det(\mathcal{D}_x) = 0\} = \{x \in \mathbb{C} \colon \det(D - x\,\mathrm{I}_{\mathfrak{d}}) = 0\}$ is a finite set. Combining this and (6.326) with the fact that the function

$$\mathbb{C} \ni x \mapsto \det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} - \det(A\mathcal{D}_x - BC) \in \mathbb{C} \tag{6.327}$$

is continuous shows that for all $x \in \mathbb{C}$ it holds that

$$\det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} - \det(A\mathcal{D}_x - BC) = 0. \tag{6.328}$$

Hence, we obtain for all $x \in \mathbb{C}$ that

$$\det\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} = \det(A\mathcal{D}_x - BC). \tag{6.329}$$

This establishes that

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det\begin{pmatrix} A & B \\ C & \mathcal{D}_0 \end{pmatrix} = \det(A\mathcal{D}_0 - BC) = \det(A\mathcal{D}_0 - BC). \tag{6.330}$$

The proof of Lemma 6.3.25 is thus completed. □

We are now in the position to formulate and prove the promised error analysis for the momentum GD optimization method in the case of the considered class of quadratic objective functions; see Proposition 6.3.26 below.

Proposition 6.3.26 (Error analysis for the momentum GD optimization method in the case of quadratic objective functions)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\kappa, \mathcal{K}, \lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}} \in (0, \infty)$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$ and $\mathcal{K} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\}$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \tfrac{1}{2}\left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2\right], \tag{6.331}$$

and let $\Theta \colon \mathbb{N}_0 \cup \{-1\} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_{-1} = \Theta_0 = \xi$ and

$$\Theta_n = \Theta_{n-1} - \tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}(\nabla\mathscr{L})(\Theta_{n-1}) + \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2(\Theta_{n-1} - \Theta_{n-2}). \tag{6.332}$$

Then

(i) it holds that $\Theta|_{\mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ is the momentum GD process for the objective function $\mathscr{L}$ with learning rates $\mathbb{N} \ni n \mapsto \frac{1}{\sqrt{\mathcal{K}\kappa}} \in [0, \infty)$, momentum decay factors $\mathbb{N} \ni n \mapsto \left[\frac{\mathcal{K}^{1/2}-\kappa^{1/2}}{\mathcal{K}^{1/2}+\kappa^{1/2}}\right]^2 \in [0, 1]$, and initial value ξ and

(ii) for every $\varepsilon \in (0, \infty)$ there exists $\mathfrak{c} \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \mathfrak{c}\left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}} + \varepsilon\right]^n \tag{6.333}$$

(cf. Definitions 3.3.4 and 6.3.1).

Proof of Proposition 6.3.26. Throughout this proof, let $\varepsilon \in (0, \infty)$, let $|||\cdot||| \colon \mathbb{R}^{(2\mathfrak{d})\times(2\mathfrak{d})} \to [0, \infty)$ satisfy for all $B \in \mathbb{R}^{(2\mathfrak{d})\times(2\mathfrak{d})}$ that

$$|||B||| = \sup_{v \in \mathbb{R}^{2\mathfrak{d}} \setminus \{0\}}\left[\frac{\|Bv\|_2}{\|v\|_2}\right], \tag{6.334}$$

let $\Theta^{(i)}\colon \mathbb{N}_0 \to \mathbb{R}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$, satisfy for all $n \in \mathbb{N}_0$ that $\Theta_n = (\Theta_n^{(1)}, \ldots, \Theta_n^{(\mathfrak{d})})$, let $\mathbf{m}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that

$$\mathbf{m}_n = -\sqrt{\mathcal{K}\kappa}(\Theta_n - \Theta_{n-1}), \tag{6.335}$$

let $\varrho \in (0, \infty)$, $\alpha \in [0, 1)$ be given by

$$\varrho = \tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} \qquad \text{and} \qquad \alpha = \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2, \tag{6.336}$$

let $M \in \mathbb{R}^{\mathfrak{d}\times\mathfrak{d}}$ be the diagonal $(\mathfrak{d} \times \mathfrak{d})$-matrix given by

$$M = \begin{pmatrix} (1 - \varrho\lambda_1 + \alpha) & & 0 \\ & \ddots & \\ 0 & & (1 - \varrho\lambda_{\mathfrak{d}} + \alpha) \end{pmatrix}, \tag{6.337}$$

let $A \in \mathbb{R}^{2\mathfrak{d}\times 2\mathfrak{d}}$ be the $((2\mathfrak{d}) \times (2\mathfrak{d}))$-matrix given by

$$A = \begin{pmatrix} M & (-\alpha\, \mathrm{I}_{\mathfrak{d}}) \\ \mathrm{I}_{\mathfrak{d}} & 0 \end{pmatrix}, \tag{6.338}$$

and let $\mathcal{S} \subseteq \mathbb{C}$ be the set given by

$$\mathcal{S} = \{\mu \in \mathbb{C}\colon (\exists\, v \in \mathbb{C}^{2\mathfrak{d}}\backslash\{0\}\colon Av = \mu v)\} = \{\mu \in \mathbb{C}\colon \det(A - \mu\, \mathrm{I}_{2\mathfrak{d}}) = 0\} \tag{6.339}$$

(cf. Definition 1.5.5). Observe that (6.332), (6.335), and the fact that

$$\begin{aligned} \tfrac{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2-(\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{4} &= \tfrac{1}{4}\Big[(\sqrt{\mathcal{K}} + \sqrt{\kappa} + \sqrt{\mathcal{K}} - \sqrt{\kappa})(\sqrt{\mathcal{K}} + \sqrt{\kappa} - [\sqrt{\mathcal{K}} - \sqrt{\kappa}])\Big] \\ &= \tfrac{1}{4}\Big[(2\sqrt{\mathcal{K}})(2\sqrt{\kappa})\Big] = \sqrt{\mathcal{K}\kappa} \end{aligned} \tag{6.340}$$

assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= -\sqrt{\mathcal{K}\kappa}(\Theta_n - \Theta_{n-1}) \\ &= -\sqrt{\mathcal{K}\kappa}\left(\Theta_{n-1} - \left[\tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right](\nabla\mathscr{L})(\Theta_{n-1}) + \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2(\Theta_{n-1} - \Theta_{n-2}) - \Theta_{n-1}\right) \\ &= \sqrt{\mathcal{K}\kappa}\left(\left[\tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right](\nabla\mathscr{L})(\Theta_{n-1}) - \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2(\Theta_{n-1} - \Theta_{n-2})\right) \\ &= \tfrac{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2-(\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{4}\left[\tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right](\nabla\mathscr{L})(\Theta_{n-1}) \\ &\quad - \sqrt{\mathcal{K}\kappa}\left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2(\Theta_{n-1} - \Theta_{n-2}) \\ &= \left[1 - \tfrac{(\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right](\nabla\mathscr{L})(\Theta_{n-1}) + \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2\left[-\sqrt{\mathcal{K}\kappa}(\Theta_{n-1} - \Theta_{n-2})\right] \\ &= \left[1 - \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2\right](\nabla\mathscr{L})(\Theta_{n-1}) + \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2\mathbf{m}_{n-1}. \end{aligned} \tag{6.341}$$

Moreover, note that (6.335) implies that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}\Theta_n &= \Theta_{n-1} + (\Theta_n - \Theta_{n-1}) \\ &= \Theta_{n-1} - \tfrac{1}{\sqrt{\mathcal{K}\kappa}}\Big(\Big[-\sqrt{\mathcal{K}\kappa}\Big](\Theta_n - \Theta_{n-1})\Big) = \Theta_{n-1} - \tfrac{1}{\sqrt{\mathcal{K}\kappa}}\mathbf{m}_n.\end{aligned} \tag{6.342}$$

In addition, observe that the assumption that $\Theta_{-1} = \Theta_0 = \xi$ and (6.335) ensure that

$$\mathbf{m}_0 = -\sqrt{\mathcal{K}\kappa}\,(\Theta_0 - \Theta_{-1}) = 0. \tag{6.343}$$

Combining this and the assumption that $\Theta_0 = \xi$ with (6.341) and (6.342) proves item (i). It thus remains to prove item (ii). For this observe that (6.331) implies that for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\big(\tfrac{\partial \mathscr{L}}{\partial \theta_i}\big)(\theta) = \lambda_i(\theta_i - \vartheta_i). \tag{6.344}$$

This, (6.332), and (6.336) imply that for all $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned}\Theta_n^{(i)} - \vartheta_i &= \Theta_{n-1}^{(i)} - \varrho\big(\tfrac{\partial \mathscr{L}}{\partial \theta_i}\big)(\Theta_{n-1}) + \alpha(\Theta_{n-1}^{(i)} - \Theta_{n-2}^{(i)}) - \vartheta_i \\ &= (\Theta_{n-1}^{(i)} - \vartheta_i) - \varrho\lambda_i(\Theta_{n-1}^{(i)} - \vartheta_i) + \alpha\big((\Theta_{n-1}^{(i)} - \vartheta_i) - (\Theta_{n-2}^{(i)} - \vartheta_i)\big) \\ &= (1 - \varrho\lambda_i + \alpha)(\Theta_{n-1}^{(i)} - \vartheta_i) - \alpha(\Theta_{n-2}^{(i)} - \vartheta_i).\end{aligned} \tag{6.345}$$

Combining this with (6.337) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\mathbb{R}^{\mathfrak{d}} \ni (\Theta_n - \vartheta) &= M(\Theta_{n-1} - \vartheta) - \alpha(\Theta_{n-2} - \vartheta) \\ &= \underbrace{\begin{pmatrix} M & (-\alpha\,\mathrm{I}_{\mathfrak{d}}) \end{pmatrix}}_{\in \mathbb{R}^{\mathfrak{d} \times 2\mathfrak{d}}} \underbrace{\begin{pmatrix} \Theta_{n-1} - \vartheta \\ \Theta_{n-2} - \vartheta \end{pmatrix}}_{\in \mathbb{R}^{2\mathfrak{d}}}.\end{aligned} \tag{6.346}$$

This and (6.338) assure that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{R}^{2\mathfrak{d}} \ni \begin{pmatrix} \Theta_n - \vartheta \\ \Theta_{n-1} - \vartheta \end{pmatrix} = \begin{pmatrix} M & (-\alpha\,\mathrm{I}_{\mathfrak{d}}) \\ \mathrm{I}_{\mathfrak{d}} & 0 \end{pmatrix} \begin{pmatrix} \Theta_{n-1} - \vartheta \\ \Theta_{n-2} - \vartheta \end{pmatrix} = A \begin{pmatrix} \Theta_{n-1} - \vartheta \\ \Theta_{n-2} - \vartheta \end{pmatrix}. \tag{6.347}$$

Induction hence proves that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{R}^{2\mathfrak{d}} \ni \begin{pmatrix} \Theta_n - \vartheta \\ \Theta_{n-1} - \vartheta \end{pmatrix} = A^n \begin{pmatrix} \Theta_0 - \vartheta \\ \Theta_{-1} - \vartheta \end{pmatrix} = A^n \begin{pmatrix} \xi - \vartheta \\ \xi - \vartheta \end{pmatrix}. \tag{6.348}$$

This implies that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
\|\Theta_n - \vartheta\|_2 &\leq \sqrt{\|\Theta_n - \vartheta\|_2^2 + \|\Theta_{n-1} - \vartheta\|_2^2} \\
&= \left\| \begin{pmatrix} \Theta_n - \vartheta \\ \Theta_{n-1} - \vartheta \end{pmatrix} \right\|_2 \\
&= \left\| A^n \begin{pmatrix} \xi - \vartheta \\ \xi - \vartheta \end{pmatrix} \right\|_2 \\
&\leq \|\!|\!| A^n |\!|\!| \left\| \begin{pmatrix} \xi - \vartheta \\ \xi - \vartheta \end{pmatrix} \right\|_2 \\
&= \|\!|\!| A^n |\!|\!| \sqrt{\|\xi - \vartheta\|_2^2 + \|\xi - \vartheta\|_2^2} \\
&= \|\!|\!| A^n |\!|\!| \sqrt{2} \|\xi - \vartheta\|_2.
\end{aligned} \tag{6.349}$$

Next note that (6.339) and Lemma 6.3.24 demonstrate that

$$\limsup_{n\to\infty} \left(\left[\|\!|\!| A^n |\!|\!| \right]^{1/n} \right) = \liminf_{n\to\infty} \left(\left[\|\!|\!| A^n |\!|\!| \right]^{1/n} \right) = \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|. \tag{6.350}$$

This implies that there exists $m \in \mathbb{N}$ which satisfies for all $n \in \mathbb{N} \cap [m, \infty)$ that

$$\left[\|\!|\!| A^n |\!|\!| \right]^{1/n} \leq \varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|. \tag{6.351}$$

Note that (6.351) implies that for all $n \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\|\!|\!| A^n |\!|\!| \leq \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n. \tag{6.352}$$

Furthermore, note that for all $n \in \mathbb{N} \cap [0, m)$ it holds that

$$\begin{aligned}
\|\!|\!| A^n |\!|\!| &= \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \left[\tfrac{\|\!|\!| A^n |\!|\!|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^n} \right] \\
&\leq \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \left[\max\left(\left\{ \tfrac{\|\!|\!| A^k |\!|\!|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^k} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right].
\end{aligned} \tag{6.353}$$

Combining this and (6.352) proves that for all $n \in \mathbb{N}_0$ it holds that

$$\|\!|\!| A^n |\!|\!| \leq \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \left[\max\left(\left\{ \tfrac{\|\!|\!| A^k |\!|\!|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^k} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right]. \tag{6.354}$$

Next observe that Lemma 6.3.25, (6.338), and the fact that for all $\mu \in \mathbb{C}$ it holds that $\mathrm{I}_\mathfrak{d}(-\mu\,\mathrm{I}_\mathfrak{d}) = -\mu\,\mathrm{I}_\mathfrak{d} = (-\mu\,\mathrm{I}_\mathfrak{d})\,\mathrm{I}_\mathfrak{d}$ ensure that for all $\mu \in \mathbb{C}$ it holds that

$$\begin{aligned}
\det(A - \mu\,\mathrm{I}_{2\mathfrak{d}}) &= \det\begin{pmatrix} (M - \mu\,\mathrm{I}_\mathfrak{d}) & (-\alpha\,\mathrm{I}_\mathfrak{d}) \\ \mathrm{I}_\mathfrak{d} & -\mu\,\mathrm{I}_\mathfrak{d} \end{pmatrix} \\
&= \det\big((M - \mu\,\mathrm{I}_\mathfrak{d})(-\mu\,\mathrm{I}_\mathfrak{d}) - (-\alpha\,\mathrm{I}_\mathfrak{d})\,\mathrm{I}_\mathfrak{d}\big) \\
&= \det\big((M - \mu\,\mathrm{I}_\mathfrak{d})(-\mu\,\mathrm{I}_\mathfrak{d}) + \alpha\,\mathrm{I}_\mathfrak{d}\big).
\end{aligned} \tag{6.355}$$

This and (6.337) demonstrate that for all $\mu \in \mathbb{C}$ it holds that

$$\begin{aligned}\det(A - \mu \operatorname{I}_{2\mathfrak{d}}) &= \det\begin{pmatrix} \big((1-\varrho\lambda_1+\alpha-\mu)(-\mu)+\alpha\big) & & 0 \\ & \ddots & \\ 0 & & \big((1-\varrho\lambda_{\mathfrak{d}}+\alpha-\mu)(-\mu)+\alpha\big) \end{pmatrix} \\ &= \prod_{i=1}^{\mathfrak{d}} \big((1-\varrho\lambda_i+\alpha-\mu)(-\mu)+\alpha\big) \\ &= \prod_{i=1}^{\mathfrak{d}} \big(\mu^2 - (1-\varrho\lambda_i+\alpha)\mu+\alpha\big).\end{aligned} \tag{6.356}$$

Moreover, note that for all $\mu \in \mathbb{C}$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned}\mu^2 - (1-\varrho\lambda_i+\alpha)\mu+\alpha &= \mu^2 - 2\mu\left[\tfrac{(1-\varrho\lambda_i+\alpha)}{2}\right] + \left[\tfrac{(1-\varrho\lambda_i+\alpha)}{2}\right]^2 + \alpha - \left[\tfrac{(1-\varrho\lambda_i+\alpha)}{2}\right]^2 \\ &= \left[\mu - \tfrac{(1-\varrho\lambda_i+\alpha)}{2}\right]^2 + \alpha - \tfrac{1}{4}[1-\varrho\lambda_i+\alpha]^2 \\ &= \left[\mu - \tfrac{(1-\varrho\lambda_i+\alpha)}{2}\right]^2 - \tfrac{1}{4}\Big[\big[1-\varrho\lambda_i+\alpha\big]^2 - 4\alpha\Big].\end{aligned} \tag{6.357}$$

Hence, we obtain that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned}&\big\{\mu \in \mathbb{C}\colon \mu^2 - (1-\varrho\lambda_i+\alpha)\mu+\alpha = 0\big\} \\ &= \left\{\mu \in \mathbb{C}\colon \left[\mu - \tfrac{(1-\varrho\lambda_i+\alpha)}{2}\right]^2 = \tfrac{1}{4}\Big[\big[1-\varrho\lambda_i+\alpha\big]^2 - 4\alpha\Big]\right\} \\ &= \left\{\tfrac{(1-\varrho\lambda_i+\alpha)+\sqrt{[1-\varrho\lambda_i+\alpha]^2-4\alpha}}{2}, \tfrac{(1-\varrho\lambda_i+\alpha)-\sqrt{[1-\varrho\lambda_i+\alpha]^2-4\alpha}}{2}, \right\} \\ &= \bigcup_{s\in\{-1,1\}} \left\{\tfrac{1}{2}\left[1-\varrho\lambda_i+\alpha+s\sqrt{(1-\varrho\lambda_i+\alpha)^2-4\alpha}\right]\right\}.\end{aligned} \tag{6.358}$$

Combining this, (6.339), and (6.356) demonstrates that

$$\begin{aligned}\mathcal{S} &= \{\mu \in \mathbb{C}\colon \det(A - \mu \operatorname{I}_{2\mathfrak{d}}) = 0\} \\ &= \left\{\mu \in \mathbb{C}\colon \left[\prod_{i=1}^{\mathfrak{d}} \big(\mu^2 - (1-\varrho\lambda_i+\alpha)\mu+\alpha\big) = 0\right]\right\} \\ &= \bigcup_{i=1}^{\mathfrak{d}} \big\{\mu \in \mathbb{C}\colon \mu^2 - (1-\varrho\lambda_i+\alpha)\mu+\alpha = 0\big\} \\ &= \bigcup_{i=1}^{\mathfrak{d}} \bigcup_{s\in\{-1,1\}} \left\{\tfrac{1}{2}\left[1-\varrho\lambda_i+\alpha+s\sqrt{(1-\varrho\lambda_i+\alpha)^2-4\alpha}\right]\right\}.\end{aligned} \tag{6.359}$$

Moreover, observe that the fact that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that $\lambda_i \geq \kappa$ and (6.336) ensure that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} 1 - \varrho\lambda_i + \alpha &\leq 1 - \varrho\kappa + \alpha = 1 - \left[\tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right]\kappa + \tfrac{(\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} \\ &= \tfrac{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2 - 4\kappa + (\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} = \tfrac{\mathcal{K}+2\sqrt{\mathcal{K}}\sqrt{\kappa}+\kappa-4\kappa+\mathcal{K}-2\sqrt{\mathcal{K}}\sqrt{\kappa}+\kappa}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} \\ &= \tfrac{2\mathcal{K}-2\kappa}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} = \tfrac{2(\sqrt{\mathcal{K}}-\sqrt{\kappa})(\sqrt{\mathcal{K}}+\sqrt{\kappa})}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} = 2\left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right] \geq 0. \end{aligned} \tag{6.360}$$

In addition, note that the fact that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that $\lambda_i \leq \mathcal{K}$ and (6.336) assure that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} 1 - \varrho\lambda_i + \alpha &\geq 1 - \varrho\mathcal{K} + \alpha = 1 - \left[\tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right]\mathcal{K} + \tfrac{(\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} \\ &= \tfrac{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2 - 4\mathcal{K} + (\sqrt{\mathcal{K}}-\sqrt{\kappa})^2}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} = \tfrac{\mathcal{K}+2\sqrt{\mathcal{K}}\sqrt{\kappa}+\kappa-4\mathcal{K}+\mathcal{K}-2\sqrt{\mathcal{K}}\sqrt{\kappa}+\kappa}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} \\ &= \tfrac{-2\mathcal{K}+2\kappa}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2} = -2\left[\tfrac{\mathcal{K}-\kappa}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right] = -2\left[\tfrac{(\sqrt{\mathcal{K}}-\sqrt{\kappa})(\sqrt{\mathcal{K}}+\sqrt{\kappa})}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}\right] \\ &= -2\left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right] \leq 0. \end{aligned} \tag{6.361}$$

Combining this, (6.360), and (6.336) implies that for all $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that

$$(1 - \varrho\lambda_i + \alpha)^2 \leq \left[2\left(\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right)\right]^2 = 4\left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2 = 4\alpha. \tag{6.362}$$

This and (6.359) demonstrate that

$$\begin{aligned} \max_{\mu\in\mathcal{S}\cup\{0\}} |\mu| &= \max_{\mu\in\mathcal{S}} |\mu| \\ &= \max_{i\in\{1,2,\ldots,d\}} \max_{s\in\{-1,1\}} \left|\frac{1}{2}\left[1 - \varrho\lambda_i + \alpha + s\sqrt{(1 - \varrho\lambda_i + \alpha)^2 - 4\alpha}\right]\right| \\ &= \frac{1}{2}\left[\max_{i\in\{1,2,\ldots,d\}} \max_{s\in\{-1,1\}} \left|\left[1 - \varrho\lambda_i + \alpha + s\sqrt{(-1)(4\alpha - [1 - \varrho\lambda_i + \alpha]^2)}\right]\right|\right] \\ &= \frac{1}{2}\left[\max_{i\in\{1,2,\ldots,d\}} \max_{s\in\{-1,1\}} \left|\left[1 - \varrho\lambda_i + \alpha + s\mathbf{i}\sqrt{4\alpha - (1 - \varrho\lambda_i + \alpha)^2}\right]\right|^2\right]^{1/2}. \end{aligned} \tag{6.363}$$

Combining this with (6.362) proves that

$$\begin{aligned} \max_{\mu\in\mathcal{S}\cup\{0\}} |\mu| &= \tfrac{1}{2}\left[\max_{i\in\{1,2,\ldots,d\}} \max_{s\in\{-1,1\}} \left(\left|1 - \varrho\lambda_i + \alpha\right|^2 + \left|s\sqrt{4\alpha - (1 - \varrho\lambda_i + \alpha)^2}\right|^2\right)\right]^{1/2} \\ &= \tfrac{1}{2}\left[\max_{i\in\{1,2,\ldots,d\}} \max_{s\in\{-1,1\}} \left((1 - \varrho\lambda_i + \alpha)^2 + 4\alpha - (1 - \varrho\lambda_i + \alpha)^2\right)\right]^{1/2} \\ &= \tfrac{1}{2}[4\alpha]^{1/2} = \sqrt{\alpha}. \end{aligned} \tag{6.364}$$

Combining (6.349) and (6.354) hence ensures that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
\left\|\Theta_n - \vartheta\right\|_2 &\leq \sqrt{2}\,\|\xi - \vartheta\|_2 |\!|\!| A^n |\!|\!| \\
&\leq \sqrt{2}\,\|\xi - \vartheta\|_2 \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|\right]^n \\
&\quad \cdot \left[\max\left(\left\{ \tfrac{|\!|\!| A^k |\!|\!|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^k} \in \mathbb{R} \colon k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\}\right)\right] \\
&= \sqrt{2}\,\|\xi - \vartheta\|_2 \left[\varepsilon + \alpha^{1/2}\right]^n \left[\max\left(\left\{ \tfrac{|\!|\!| A^k |\!|\!|}{(\varepsilon + \alpha^{1/2})^k} \in \mathbb{R} \colon k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\}\right)\right] \\
&= \sqrt{2}\,\|\xi - \vartheta\|_2 \left[\varepsilon + \tfrac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}}\right]^n \left[\max\left(\left\{ \tfrac{|\!|\!| A^k |\!|\!|}{(\varepsilon + \alpha^{1/2})^k} \in \mathbb{R} \colon k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\}\right)\right].
\end{aligned} \tag{6.365}$$

This establishes item (ii). The proof of Proposition 6.3.26 it thus completed. □

6.3.5.3 Comparison of the convergence speeds of GD optimization with and without momentum

In this subsection we provide in Corollary 6.3.28 below a comparison between the convergence speeds of the plain-vanilla GD optimization method and the momentum GD optimization method. Our proof of Corollary 6.3.28 employs the auxiliary and elementary estimate in Lemma 6.3.27 below, the refined error analysis for the plain-vanilla GD optimization method in Section 6.3.5.1 above (see Lemma 6.3.22 and Lemma 6.3.23 in Section 6.3.5.1), as well as the error analysis for the momentum GD optimization method in Section 6.3.5.2 above (see Proposition 6.3.26 in Section 6.3.5.2).

Lemma 6.3.27 (Comparison of the convergence rates of the GD optimization method and the momentum GD optimization method)**.** *Let $\mathcal{K}, \kappa \in (0, \infty)$ satisfy $\kappa < \mathcal{K}$. Then*

$$\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} < \frac{\mathcal{K} - \kappa}{\mathcal{K} + \kappa}. \tag{6.366}$$

Proof of Lemma 6.3.27. Note that the fact that $\mathcal{K} - \kappa > 0 < 2\sqrt{\mathcal{K}}\sqrt{\kappa}$ ensures that

$$\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} = \frac{(\sqrt{\mathcal{K}} - \sqrt{\kappa})(\sqrt{\mathcal{K}} + \sqrt{\kappa})}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} = \frac{\mathcal{K} - \kappa}{\mathcal{K} + 2\sqrt{\mathcal{K}}\sqrt{\kappa} + \kappa} < \frac{\mathcal{K} - \kappa}{\mathcal{K} + \kappa}. \tag{6.367}$$

The proof of Lemma 6.3.27 it thus completed. □

Corollary 6.3.28 (Convergence speed comparisons between the GD optimization method and the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\kappa, \mathcal{K}, \lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}} \in (0, \infty)$, $\xi = (\xi_1, \ldots, \xi_{\mathfrak{d}})$, $\vartheta = (\vartheta_1, \ldots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\} <$*

$\max\{\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}}\} = \mathcal{K}$, *let* $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ *that*

$$\mathscr{L}(\theta) = \tfrac{1}{2}\left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2\right], \tag{6.368}$$

for every $\gamma \in (0, \infty)$ *let* $\Theta^\gamma \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that*

$$\Theta_0^\gamma = \xi \qquad \text{and} \qquad \Theta_n^\gamma = \Theta_{n-1}^\gamma - \gamma(\nabla \mathscr{L})(\Theta_{n-1}^\gamma), \tag{6.369}$$

and let $\mathcal{M} \colon \mathbb{N}_0 \cup \{-1\} \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that* $\mathcal{M}_{-1} = \mathcal{M}_0 = \xi$ *and*

$$\mathcal{M}_n = \mathcal{M}_{n-1} - \tfrac{4}{(\sqrt{\mathcal{K}}+\sqrt{\kappa})^2}(\nabla \mathscr{L})(\mathcal{M}_{n-1}) + \left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}}\right]^2 (\mathcal{M}_{n-1} - \mathcal{M}_{n-2}). \tag{6.370}$$

Then

(i) there exist $\gamma, \mathfrak{c} \in (0, \infty)$ *such that for all* $n \in \mathbb{N}_0$ *it holds that*

$$\|\Theta_n^\gamma - \vartheta\|_2 \leq \mathfrak{c}\left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n, \tag{6.371}$$

(ii) it holds for all $\gamma \in (0, \infty), n \in \mathbb{N}_0$ *that*

$$\|\Theta_n^\gamma - \vartheta\|_2 \geq \left[\min\{|\xi_1 - \vartheta_1|, \ldots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}\right]\left[\tfrac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n, \tag{6.372}$$

(iii) for every $\varepsilon \in (0, \infty)$ *there exists* $\mathfrak{c} \in (0, \infty)$ *such that for all* $n \in \mathbb{N}_0$ *it holds that*

$$\|\mathcal{M}_n - \vartheta\|_2 \leq \mathfrak{c}\left[\tfrac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}} + \varepsilon\right]^n, \tag{6.373}$$

and

(iv) it holds that $\frac{\sqrt{\mathcal{K}}-\sqrt{\kappa}}{\sqrt{\mathcal{K}}+\sqrt{\kappa}} < \frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}$

(cf. Definition 3.3.4).

Proof of Corollary 6.3.28. First, note that Lemma 6.3.22 proves item (i). Next observe that Lemma 6.3.23 establishes item (ii). In addition, note that Proposition 6.3.26 proves item (iii). Finally, observe that Lemma 6.3.27 establishes item (iv). The proof of Corollary 6.3.28 is thus complete. □

Corollary 6.3.28 above, roughly speaking, shows in the case of the considered class of quadratic objective functions that the momentum GD optimization method in (6.370) outperforms the classical plain-vanilla GD optimization method (and, in particular, the classical plain-vanilla GD optimization method in (6.296) in Lemma 6.3.22 above) provided

that the parameters $\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}} \in (0, \infty)$ in the objective function in (6.368) satisfy the assumption that

$$\min\{\lambda_1, \ldots, \lambda_{\mathfrak{d}}\} < \max\{\lambda_1, \ldots, \lambda_{\mathfrak{d}}\}. \tag{6.374}$$

The next elementary result, Lemma 6.3.29 below, demonstrates that the momentum GD optimization method in (6.370) and the plain-vanilla GD optimization method in (6.296) in Lemma 6.3.22 above coincide in the case where $\min\{\lambda_1, \ldots, \lambda_{\mathfrak{d}}\} = \max\{\lambda_1, \ldots, \lambda_{\mathfrak{d}}\}$.

Lemma 6.3.29 (Concurrence of the GD optimization method and the momentum GD optimization method)**.** *Let $\mathfrak{d} \in \mathbb{N}$, $\xi, \vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\alpha \in (0, \infty)$, let $\mathscr{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathscr{L}(\theta) = \tfrac{\alpha}{2} \|\theta - \vartheta\|_2^2, \tag{6.375}$$

let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{2}{(\alpha+\alpha)} (\nabla \mathscr{L})(\Theta_{n-1}), \tag{6.376}$$

and let $\mathcal{M} \colon \mathbb{N}_0 \cup \{-1\} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\mathcal{M}_{-1} = \mathcal{M}_0 = \xi$ and

$$\mathcal{M}_n = \mathcal{M}_{n-1} - \tfrac{4}{(\sqrt{\alpha}+\sqrt{\alpha})^2} (\nabla \mathscr{L})(\mathcal{M}_{n-1}) + \left[\tfrac{\sqrt{\alpha}-\sqrt{\alpha}}{\sqrt{\alpha}+\sqrt{\alpha}}\right]^2 (\mathcal{M}_{n-1} - \mathcal{M}_{n-2}) \tag{6.377}$$

(cf. Definition 3.3.4). Then

(i) it holds that $\mathcal{M}|_{\mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ is the momentum GD process for the objective function $\mathscr{L}$ with learning rates $\mathbb{N} \ni n \mapsto 1/\alpha \in [0, \infty)$, momentum decay factors $\mathbb{N} \ni n \mapsto 0 \in [0, 1]$, and initial value ξ,

(ii) it holds for all $n \in \mathbb{N}_0$ that $\mathcal{M}_n = \Theta_n$, and

(iii) it holds for all $n \in \mathbb{N}$ that $\Theta_n = \vartheta = \mathcal{M}_n$

(cf. Definition 6.3.1).

Proof of Lemma 6.3.29. First, note that (6.377) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathcal{M}_n = \mathcal{M}_{n-1} - \tfrac{4}{(2\sqrt{\alpha})^2} (\nabla \mathscr{L})(\mathcal{M}_{n-1}) = \mathcal{M}_{n-1} - \tfrac{1}{\alpha} (\nabla \mathscr{L})(\mathcal{M}_{n-1}). \tag{6.378}$$

Combining this with the assumption that $\mathcal{M}_0 = \xi$ establishes item (i). Next note that (6.376) ensures that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \tfrac{1}{\alpha} (\nabla \mathscr{L})(\Theta_{n-1}). \tag{6.379}$$

Combining this with (6.378) and the assumption that $\Theta_0 = \xi = \mathcal{M}_0$ proves item (ii). Furthermore, observe that Lemma 5.8.4 assures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$(\nabla \mathscr{L})(\theta) = \tfrac{\alpha}{2}(2(\theta - \vartheta)) = \alpha(\theta - \vartheta). \tag{6.380}$$

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

$$\Theta_n = \vartheta. \tag{6.381}$$

We now prove (6.381) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (6.379) and (6.380) imply that

$$\Theta_1 = \Theta_0 - \tfrac{1}{\alpha}(\nabla \mathscr{L})(\Theta_0) = \xi - \tfrac{1}{\alpha}(\alpha(\xi - \vartheta)) = \xi - (\xi - \vartheta) = \vartheta. \tag{6.382}$$

This establishes (6.381) in the base case $n = 1$. For the induction step observe that (6.379) and (6.380) assure that for all $n \in \mathbb{N}$ with $\Theta_n = \vartheta$ it holds that

$$\Theta_{n+1} = \Theta_n - \tfrac{1}{\alpha}(\nabla \mathscr{L})(\Theta_n) = \vartheta - \tfrac{1}{\alpha}(\alpha(\vartheta - \vartheta)) = \vartheta. \tag{6.383}$$

Induction thus proves (6.381). Combining (6.381) and item (ii) establishes item (iii). The proof of Lemma 6.3.29 is thus complete. □

6.3.6 Numerical comparisons for GD optimization with and without momentum

In this subsection we provide in Example 6.3.30, Source code 6.1, and Figure 6.1 a numerical comparison of the plain-vanilla GD optimization method and the momentum GD optimization method in the case of the specific quadratic optimization problem in (6.384)–(6.385) below.

Example 6.3.30. *Let $\mathcal{K} = 10$, $\kappa = 1$, $\vartheta = (\vartheta_1, \vartheta_2) \in \mathbb{R}^2$, $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$ satisfy*

$$\vartheta = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \tag{6.384}$$

let $\mathscr{L} \colon \mathbb{R}^2 \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ that

$$\mathscr{L}(\theta) = \left(\tfrac{\kappa}{2}\right)|\theta_1 - \vartheta_1|^2 + \left(\tfrac{\mathcal{K}}{2}\right)|\theta_2 - \vartheta_2|^2, \tag{6.385}$$

let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0 = \xi$ and

$$\begin{aligned} \Theta_n &= \Theta_{n-1} - \tfrac{2}{\mathcal{K}+\kappa}(\nabla \mathscr{L})(\Theta_{n-1}) = \Theta_{n-1} - \tfrac{2}{11}(\nabla \mathscr{L})(\Theta_{n-1}) \\ &= \Theta_{n-1} - 0.\overline{18}(\nabla \mathscr{L})(\Theta_{n-1}) \approx \Theta_{n-1} - 0.18(\nabla \mathscr{L})(\Theta_{n-1}), \end{aligned} \tag{6.386}$$

and let $\mathcal{M}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *and* $\mathbf{m}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $n \in \mathbb{N}$ *that* $\mathcal{M}_0 = \xi$, $\mathbf{m}_0 = 0$, $\mathcal{M}_n = \mathcal{M}_{n-1} - 0.3\,\mathbf{m}_n$, *and*

$$\begin{aligned} \mathbf{m}_n &= 0.5\,\mathbf{m}_{n-1} + (1-0.5)(\nabla\mathscr{L})(\mathcal{M}_{n-1}) \\ &= 0.5\,(\mathbf{m}_{n-1} + (\nabla\mathscr{L})(\mathcal{M}_{n-1})). \end{aligned} \tag{6.387}$$

Then

(i) it holds for all $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ *that*

$$(\nabla\mathscr{L})(\theta) = \begin{pmatrix} \kappa(\theta_1 - \vartheta_1) \\ \mathcal{K}(\theta_2 - \vartheta_2) \end{pmatrix} = \begin{pmatrix} \theta_1 - 1 \\ 10\,(\theta_2 - 1) \end{pmatrix}, \tag{6.388}$$

(ii) it holds that

$$\Theta_0 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \tag{6.389}$$

$$\begin{aligned} \Theta_1 &= \Theta_0 - \tfrac{2}{11}(\nabla\mathscr{L})(\Theta_0) \approx \Theta_0 - 0.18(\nabla\mathscr{L})(\Theta_0) \\ &= \begin{pmatrix} 5 \\ 3 \end{pmatrix} - 0.18 \begin{pmatrix} 5-1 \\ 10(3-1) \end{pmatrix} = \begin{pmatrix} 5 - 0.18 \cdot 4 \\ 3 - 0.18 \cdot 10 \cdot 2 \end{pmatrix} \\ &= \begin{pmatrix} 5 - 0.72 \\ 3 - 3.6 \end{pmatrix} = \begin{pmatrix} 4.28 \\ -0.6 \end{pmatrix}, \end{aligned} \tag{6.390}$$

$$\begin{aligned} \Theta_2 &\approx \Theta_1 - 0.18(\nabla\mathscr{L})(\Theta_1) = \begin{pmatrix} 4.28 \\ -0.6 \end{pmatrix} - 0.18 \begin{pmatrix} 4.28 - 1 \\ 10(-0.6 - 1) \end{pmatrix} \\ &= \begin{pmatrix} 4.28 - 0.18 \cdot 3.28 \\ -0.6 - 0.18 \cdot 10 \cdot (-1.6) \end{pmatrix} = \begin{pmatrix} 4.10 - 0.18 \cdot 2 - 0.18 \cdot 0.28 \\ -0.6 + 1.8 \cdot 1.6 \end{pmatrix} \\ &= \begin{pmatrix} 4.10 - 0.36 - 2 \cdot 9 \cdot 4 \cdot 7 \cdot 10^{-4} \\ -0.6 + 1.6 \cdot 1.6 + 0.2 \cdot 1.6 \end{pmatrix} = \begin{pmatrix} 3.74 - 9 \cdot 56 \cdot 10^{-4} \\ -0.6 + 2.56 + 0.32 \end{pmatrix} \\ &= \begin{pmatrix} 3.74 - 504 \cdot 10^{-4} \\ 2.88 - 0.6 \end{pmatrix} = \begin{pmatrix} 3.6896 \\ 2.28 \end{pmatrix} \approx \begin{pmatrix} 3.69 \\ 2.28 \end{pmatrix}, \end{aligned} \tag{6.391}$$

$$\begin{aligned} \Theta_3 &\approx \Theta_2 - 0.18(\nabla\mathscr{L})(\Theta_2) \approx \begin{pmatrix} 3.69 \\ 2.28 \end{pmatrix} - 0.18 \begin{pmatrix} 3.69 - 1 \\ 10(2.28 - 1) \end{pmatrix} \\ &= \begin{pmatrix} 3.69 - 0.18 \cdot 2.69 \\ 2.28 - 0.18 \cdot 10 \cdot 1.28 \end{pmatrix} = \begin{pmatrix} 3.69 - 0.2 \cdot 2.69 + 0.02 \cdot 2.69 \\ 2.28 - 1.8 \cdot 1.28 \end{pmatrix} \\ &= \begin{pmatrix} 3.69 - 0.538 + 0.0538 \\ 2.28 - 1.28 - 0.8 \cdot 1.28 \end{pmatrix} = \begin{pmatrix} 3.7438 - 0.538 \\ 1 - 1.28 + 0.2 \cdot 1.28 \end{pmatrix} \\ &= \begin{pmatrix} 3.2058 \\ 0.256 - 0.280 \end{pmatrix} = \begin{pmatrix} 3.2058 \\ -0.024 \end{pmatrix} \approx \begin{pmatrix} 3.21 \\ -0.02 \end{pmatrix}, \end{aligned} \tag{6.392}$$

$$\vdots$$

and

(iii) it holds that

$$\mathcal{M}_0 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \tag{6.393}$$

$$\begin{aligned} \mathbf{m}_1 &= 0.5\,(\mathbf{m}_0 + (\nabla \mathscr{L})(\mathcal{M}_0)) = 0.5\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 5-1 \\ 10(3-1) \end{pmatrix}\right) \\ &= \begin{pmatrix} 0.5\,(0+4) \\ 0.5\,(0+10\cdot 2) \end{pmatrix} = \begin{pmatrix} 2 \\ 10 \end{pmatrix}, \end{aligned} \tag{6.394}$$

$$\mathcal{M}_1 = \mathcal{M}_0 - 0.3\,\mathbf{m}_1 = \begin{pmatrix} 5 \\ 3 \end{pmatrix} - 0.3\begin{pmatrix} 2 \\ 10 \end{pmatrix} = \begin{pmatrix} 4.4 \\ 0 \end{pmatrix}, \tag{6.395}$$

$$\begin{aligned} \mathbf{m}_2 &= 0.5\,(\mathbf{m}_1 + (\nabla \mathscr{L})(\mathcal{M}_1)) = 0.5\left(\begin{pmatrix} 2 \\ 10 \end{pmatrix} + \begin{pmatrix} 4.4-1 \\ 10(0-1) \end{pmatrix}\right) \\ &= \begin{pmatrix} 0.5\,(2+3.4) \\ 0.5\,(10-10) \end{pmatrix} = \begin{pmatrix} 2.7 \\ 0 \end{pmatrix}, \end{aligned} \tag{6.396}$$

$$\mathcal{M}_2 = \mathcal{M}_1 - 0.3\,\mathbf{m}_2 = \begin{pmatrix} 4.4 \\ 0 \end{pmatrix} - 0.3\begin{pmatrix} 2.7 \\ 0 \end{pmatrix} = \begin{pmatrix} 4.4-0.81 \\ 0 \end{pmatrix} = \begin{pmatrix} 3.59 \\ 0 \end{pmatrix}, \tag{6.397}$$

$$\begin{aligned} \mathbf{m}_3 &= 0.5\,(\mathbf{m}_2 + (\nabla \mathscr{L})(\mathcal{M}_2)) = 0.5\left(\begin{pmatrix} 2.7 \\ 0 \end{pmatrix} + \begin{pmatrix} 3.59-1 \\ 10(0-1) \end{pmatrix}\right) \\ &= \begin{pmatrix} 0.5\,(2.7+2.59) \\ 0.5\,(0-10) \end{pmatrix} = \begin{pmatrix} 0.5\cdot 5.29 \\ 0.5(-10) \end{pmatrix} \\ &= \begin{pmatrix} 2.5+0.145 \\ -5 \end{pmatrix} = \begin{pmatrix} 2.645 \\ -5 \end{pmatrix} \approx \begin{pmatrix} 2.65 \\ -5 \end{pmatrix}, \end{aligned} \tag{6.398}$$

$$\begin{aligned} \mathcal{M}_3 &= \mathcal{M}_2 - 0.3\,\mathbf{m}_3 \approx \begin{pmatrix} 3.59 \\ 0 \end{pmatrix} - 0.3\begin{pmatrix} 2.65 \\ -5 \end{pmatrix} \\ &= \begin{pmatrix} 3.59-0.795 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 3-0.205 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 2.795 \\ 1.5 \end{pmatrix} \approx \begin{pmatrix} 2.8 \\ 1.5 \end{pmatrix}, \end{aligned} \tag{6.399}$$

$$\vdots$$